Short T Dynamic Panel Data Models with Individual and Interactive Time Effects*

Kazuhiko Hayakawa
Hiroshima University

M. Hashem Pesaran
University of Southern California, USA, and Trinity College, Cambridge

L. Vanessa Smith
University of York

2 September 2018

Abstract

This paper proposes a quasi maximum likelihood estimator for short T dynamic fixed effects panel data models allowing for interactive time effects through a multi-factor error structure. The proposed estimator is robust to the heterogeneity of the initial values and common unobserved effects, whilst at the same time allowing for standard fixed and time effects. It is applicable to both stationary and unit root cases. Order conditions for identification of the number of interactive effects are established, and conditions are derived under which the parameters are almost surely locally identified. It is shown that global identification is possible only when the model does not contain lagged dependent variables. The QML estimator is proven to be consistent and asymptotically normally distributed. A sequential multiple testing likelihood ratio procedure is also proposed for estimation of the number of factors which is shown to be consistent. Finite sample results obtained from Monte Carlo simulations show that the proposed procedure for determining the number of factors performs very well and the quasi ML estimator has small bias and RMSE, and correct empirical size even when the number of factors is estimated. An empirical application, revisiting the growth convergence literature is also provided.

JEL Classifications: C12, C13, C23

Keywords: short T dynamic panels, unobserved common factors, quasi maximum likelihood, interactive time effects, multiple testing, sequential likelihood ratio tests, output and growth convergence

*The authors would like to thank Alex Chudik and Ron Smith for helpful suggestions, and Vasilis Sarafidis for helpful comments on a preliminary version of the paper. Part of this paper was written whilst Hayakawa was visiting the University of Cambridge as a JSPS Postdoctoral Fellow for Research Abroad. He acknowledges financial support from the JSPS Fellowship and the Grant-in-Aid for Scientific Research (KAKENHI 22730178, 25780153) provided by the JSPS. Pesaran and Smith acknowledge financial support from the ESRC Grant No. ES/I031626/1.

1 Introduction

There now exists an extensive literature on the estimation of linear dynamic panel data models where the time dimension (T) is short and fixed relative to the cross section dimension (N), which is large. Such panels are usually referred to as *micro panels*, and often arise in microeconometric applications. For example, many empirical applications based on survey data such as the British Household Panel Surveys (BHPS) and the Panel Study in Income Dynamics (PSID) are characterised by data covering relatively short time periods. Short T panels also arise in the cross country empirical growth literature where data is typically averaged over five to seven years to eliminate the business cycle effects. It is now quite common to include dynamics in such studies in addition to individual and time fixed effects, the former being particularly important to capture individual characteristics, and the latter to control for common shocks and the influence of aggregate trends. Empirical applications of dynamic panel data models with both individual and time effects using survey data include, for example, the studies of Guariglia and Rossi (2002) and Prior (2010). In the context of growth empirics these include Islam (1995), Caselli et al. (1996), and Aiginger and Falk (2005) among others. Although such studies feature individual and time effects along with dynamics, it is rare to find studies that allow for error cross section dependence as well. In many empirical applications time dummies are used to deal with cross section dependence, which is valid only if the time effect is homogeneous over the cross section units.

Both generalized method of moments (GMM) and likelihood approaches have been advanced to estimate such panel data models. See, for example, Anderson and Hsiao (1981), Arellano and Bond (1991), Arellano and Bover (1995), Blundell and Bond (1998), Hsiao et al. (2002), Binder et al. (2005) and Moral-Benito (2013). However, this literature assumes that the errors are cross sectionally independent, which might not hold in many applications where cross section units are subject to common unobserved effects, or possibly spatial or network spillover effects. Ignoring cross section dependence can have important consequences for conventional estimators of dynamic panels. Phillips and Sul (2007) study the impact of cross section dependence, modelled as a factor structure, on the inconsistency of the pooled least squares estimate of a short dynamic panel regression. Sarafidis and Robertson (2009) investigate the properties of a number of standard widely used GMM estimators under cross section dependence and show that such estimators are inconsistent.

In applications where the spatial patterns are important and can be characterised by known spatial weight matrices, error cross section dependence is typically modelled as spatial autoregressions and estimated jointly with the other parameters of the dynamic panel data model. Such models with short T are considered, for example, by Elhorst (2005) and Su and Yang (2015) for random effects as well as fixed effects specifications. In the latter case the first-difference operator is applied to eliminate the fixed effects and then the transformed likelihood approach of Hsiao et al. (2002) is used to deal with the initial value problem. The treatment of the initial values in spatial dynamic panel data models poses additional difficulties and requires further investigation. Jacobs et al. (2009) discuss GMM estimation of dynamic fixed effect panel data models featuring spatially correlated errors and endogenous interaction. See Lee and Yu (2010) for a review.

In addition to the spatial effects it is also likely that the error cross section dependence could be a result of omitted unobserved common factor(s). This class of models has been the subject of intensive research over the recent years and robust estimation procedures have been advanced in the case of panels where N and T are both large. See, for example, Pesaran (2006), Bai (2009), Pesaran and Tosetti (2011), Chudik et al. (2011), and Kapetanios et al. (2011). By comparison, less work has been done on estimation of short T dynamic panels where error cross section dependence is due to unobserved common factors, also known as interactive effects. An early contribution by MaCurdy (1982) features panel models with an error structure that combines factor schemes with autoregressive-moving average models estimated by maximum likelihood and used to analyse the error process associated with the earnings of prime age males. Further recent related literature will be considered in the next section. A recent survey of panel data models with error cross section dependence and short T can be found in Sarafidis and Wansbeek (2012).

Motivated by the practices and requirements of the empirical literature, in this paper we explicitly consider individual and time effects within a dynamic panel data model with short T, allowing in addition for interactive effects. In the analysis of output and growth convergence for example, accounting for interactive effects allows to capture the idea that all economies have access, possibly with different degrees, to the same pool of technological knowledge (Pesaran, 2007). Building on the work of Hsiao et al. (2002), we propose an alternative quasi maximum likelihood (QML) approach applied to the panel data model after first-differencing. In this way, we account for heterogeneity of the initial values and the common factors in an integrated framework. We establish order conditions for identification of the number of interactive effects, and derive conditions under which the parameters are almost surely locally identified. Global identification is possible only when the model does not contain lagged dependent variables. The QML estimators are shown to be consistent and asymptotically normally distributed both for stationary and unit root cases. Most importantly, for the practical implementation of our approach we propose a sequential multiple testing likelihood ratio (MTLR) procedure to estimate the number of interactive effects, which delivers a consistent estimator of the true number of factors. The proposed method can be readily extended to a panel VAR framework as in Binder et al. (2005). Monte Carlo simulations are carried out to investigate the finite sample performance of the QML estimator and the MTLR procedure, followed by an application of the approach to growth convergence.

The rest of this paper is organised as follows. Section 2 reviews the recent related literature. Section 3 sets out the dynamic panel data model and its assumptions. Section 4 develops the quasi likelihood approach and derives a solution using an eigenvalue approach. Identification of the number of factors and the parameters of the model are discussed in Section 5. Section 6 establishes the consistency of the QML estimator and derives its asymptotic distribution. Section 7 presents the sequential MTLR procedure for estimating the number of factors. Section 8 describes the Monte Carlo experiments and provides finite sample results on the performance of the sequential MTLR estimator for the number of factors, and the proposed QML estimator. An empirical application to growth convergence is provided in Section 9. The final section presents some concluding remarks. All technical proofs are provided in the Appendix. Details of alternative GMM estimators used in the Monte Carlo experiments together with additional Monte Carlo results are provided in an online supplement.

Notations: Let $\mathbf{w} = (w_1, w_2, ..., w_n)'$ and $\mathbf{A} = (a_{ij})$ be an $n \times 1$ vector and an $n \times n$ matrix, respectively. Denote the Euclidean norm of $\mathbf{w}$ and the Frobenius norm of $\mathbf{A}$ by $\|\mathbf{w}\| = (\Sigma_{i=1}^n w_i^2)^{1/2}$ and $\|\mathbf{A}\| = [Tr(\mathbf{A}'\mathbf{A})]^{1/2}$ respectively, and the largest and smallest eigenvalue of $\mathbf{A}$ by $\lambda_{max}(\mathbf{A})$ and $\lambda_{min}(\mathbf{A})$. $\boldsymbol{\tau}_T$ is a $T \times 1$ vector of ones, $\boldsymbol{\tau}_T = (1, 1, ..., 1)'$. If $\{y_n\}_{n=1}^{\infty}$ is any real sequence and $\{x_n\}_{n=1}^{\infty}$ is a sequence of positive real numbers, then $y_n = O(x_n)$ if there exists a positive finite constant C_0 such that $|y_n|/x_n \leq C_0$ for all n. $y_n = o(x_n)$ if $f_n/g_n \to 0$ as $n \to \infty$. If $\{y_n\}_{n=1}^{\infty}$ and $\{x_n\}_{n=1}^{\infty}$ are both positive sequences of real numbers, then $y_n = \ominus(x_n)$ if there exists $N_0 \geq 1$ and positive finite constants K_0 and K_1 such that $\inf_{n \geq N_0}(y_n/x_n) \geq K_0$ and $\sup_{n \geq N_0}(y_n/x_n) \leq K_1$. Positive, possibly large, fixed constants will be denoted by K (and if needed by K_0, K_1 and so on) that could take different values in different equations. Small positive constants will be denoted by ϵ. $E_0(.)$ denotes expectations taken under the true probability measure. $\to_p$ and $\overset{a.s.}{\to}$ denote convergence in probability and almost sure (a.s.) convergence, respectively. $\to_d$ denotes convergence in distribution for fixed T and as $N \to \infty$.

2 Related Literature

There is a substantial literature on estimation of short T dynamic panels. Such models are typically estimated using the generalized method of moments (GMM) applied to the first-differenced version of panel data models. The GMM approach is quite general and has been applied to a variety of dynamic panels. See, for example, Anderson and Hsiao (1981 and 1982), Holtz-Eakin et al. (1988), Arellano and Bond (1991), Ahn and Schmidt (1995), Arellano and Bover (1995), and Blundell and Bond (1998). However, these papers primarily focus on models with individual effects and when they consider time effects this is done assuming they are homogeneous across the individual units. Short T dynamic panels with heterogeneous time effects modelled as multi-factor error processes are considered by Ahn, Lee and Schmidt (2001,2013), and more recently by Bai (2013).[1] Ahn et al. (2001) consider a single factor error structure and propose a quasi-differencing approach to eliminate the factor, and then apply GMM to consistently estimate the parameters. The quasi-differencing transformation was originally proposed by Chamberlain (1984). Holtz-Eakin et al. (1988) implement it in the context of a bivariate panel autoregression. Nauges and Thomas (2003) follow the same approach in addition to prior first-differencing to eliminate the fixed effects, which they consider separately from the single factor error structure they assume for the errors. Ahn et al. (2013) extend their quasi-differencing approach to a multi-factor error structure. More recently, Hayakawa (2012) proposes a GMM estimator based on the projection method while Robertson and Sarafidis (2015) propose an instrumental variable estimation procedure that introduces new parameters to represent the unobserved covariances between the instruments and the unobserved factors. They show that the resulting estimator is asymptotically more efficient than the GMM estimator based on quasi-differencing as it exploits extra restrictions assumed. See also comments on this approach by Ahn (2015) and Hayakawa (2016).

As an alternative to GMM, Bai (2013) proposes a quasi-maximum likelihood approach applied to the original dynamic panel data model without differencing, treating time effects as free parameters, and without explicitly allowing for individual effects. To deal with

[1] Bai (2013) refers to models with multi-factor error structures as panels with interactive effects.

possible correlations between the factor loadings and the regressors Bai follows Mundlak (1978) and Chamberlain (1982) and specifies linear relationships between the factor loadings and the regressors to be estimated along with the other parameters. However, he continues to assume that all factor loadings (including the ones associated with the individual effects) are uncorrelated with the errors.

We also use a likelihood framework, but unlike Bai (2013) we allow for unrestricted individual effects possibly correlated with the errors. Our procedure also differs from the one suggested by Bai (2013) since we apply the maximum likelihood estimation to first-differences with individual effects eliminated. Our proposed estimation method can be viewed as a generalization of the transformed likelihood approach of Hsiao et al. (2002) where we now allow for unobserved common effects through the use of a multi-factor error structure. In this way we deal with error cross sectional dependence as well as the dependence of the initial values on the model parameters. Finally, we propose a sequential multiple testing likelihood procedure to consistently estimate the number of factors which is not considered by Bai (2013).

3 A dynamic panel data model with interactive error components

We begin with the following standard dynamic panel data model with time and fixed effects

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i + \delta_t + \zeta_{it}, \text{ for } t = 0, 1, 2, ..., T, \text{and } i = 1, 2, ..., N, \tag{1}$$

where $\mathbf{x}_{it}$ is a $k \times 1$ vector of regressors that vary both across i and t, $|\gamma| \leq K$, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown coefficients, with $\|\boldsymbol{\beta}\| < K$, and K denotes a finite positive constant. α_i and δ_t denote unit-specific fixed effects and time effects, respectively. We consider T to be fixed, and allow $N \to \infty$, under which the unit root case where $|\gamma| = 1$ is also covered. It is assumed that the observations $\{y_{it}, \mathbf{x}_{it}, \text{ for } t = 0, 1, ..., T; i = 1, 2, ..., N\}$ are available for estimation of γ and $\boldsymbol{\beta}$, which are the parameters of interest.

Specification (1) is the standard short T dynamic panel data model used extensively in the empirical literature assuming that the errors, ζ_{it}, are independently distributed across i and t. In this paper we contribute to this literature by allowing the errors to have the following multi-factor structure

$$\zeta_{it} = \boldsymbol{\eta}_i'\mathbf{f}_t + u_{it}, \tag{2}$$

where $\boldsymbol{\eta}_i'\mathbf{f}_t$ is an interactive effect with $\mathbf{f}_t$ an $m \times 1$ vector of unobserved common factors, $\boldsymbol{\eta}_i$ an $m \times 1$ vector of associated factor loadings, and u_{it} denotes the remaining idiosyncratic error term. The above specification contains a number of models considered in the literature and reviewed in Section 2 above as special cases. It also provides a direct generalization of Hsiao and Tahmiscioglu (2008) who consider estimation of (1) with IID errors using the transformed MLE procedure. The model considered by Ahn et al. (2013) allows for errors to have the multi-factor error structure as in (2) but does not explicitly allow for time effects in (1).

We propose an extension of the transformed MLE by treating the unknown factors as fixed parameters to be estimated for each t, but following Ahn, Lee and Schmidt (2001,2013) we assume the factor loadings to be random and distributed independently of the errors, u_{it}, and the regressors, $\mathbf{x}_{it}$. We also contribute to the analysis of identification of short T dynamic models with a multiple factor error structure, and derive order conditions for identification of m and the parameters of interest, γ and $\boldsymbol{\beta}$. Initially, we develop our proposed estimation method assuming that m is known, and consider the problem of consistent estimation of m in Section 7.1.

We make the following assumptions:

Assumption 1 *The idiosyncratic errors, u_{it}, for $i = 1, 2, ..., N$ are distributed independently across i and over t with zero means and constant variance, σ^2, such that $0 < \sigma^2 < K$, and $\sup_{i,t} E|u_{it}|^{4+\epsilon} < K$.*

Assumption 2 *The time effects, δ_t, for $t = 1, 2, ..., T$, and the $m \times 1$ vector of factors $\mathbf{f}_t$, vary across t, so that $\Delta\delta_t \neq 0$ and $\mathbf{g}_t = \Delta\mathbf{f}_t \neq \mathbf{0}$ at least for some $t = 2, ..., T$, $m < T$, and $\sup_t \|\mathbf{g}_t\| < K$.*

Assumption 3 *The regressors, $\mathbf{x}_{it}$, for $i = 1, 2, ..., N$ are distributed independently of $u_{it'}$ and $\boldsymbol{\eta}_i$, for all i, t, and t', and their first-differences, $\Delta\mathbf{x}_{it}$, follow general linear stationary time series processes*

$$\Delta\mathbf{x}_{it} = \mathbf{c}_x + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\varepsilon}_{i,t-j}, \text{ for } i = 1, 2, ..., N, \tag{3}$$

where $\mathbf{c}_x$ and $\boldsymbol{\Psi}_j$ for $j = 0, 1, ...$ are $k \times 1$ vector and $k \times k$ matrices of fixed constants such that $\|\mathbf{c}_x\| < K$, and $\sum_{j=0}^{\infty} \|\boldsymbol{\Psi}_j\| < K$. Further $\boldsymbol{\varepsilon}_{it} \sim IID(\mathbf{0}, \mathbf{I}_k)$, with $\sup_{i,t} E\|\boldsymbol{\varepsilon}_{it}\|^{4+\epsilon} < K$.

Assumption 4 *The unit specific fixed effects, α_i, for $i = 1, 2, ..., N$ are allowed to be correlated with $\mathbf{x}_{jt}$, $\boldsymbol{\eta}_j$, and u_{jt}, for all i, j and t, and could be deterministic and uniformly bounded, $\sup_i |\alpha_i| < K$, or stochastic and uniformly bounded, $\sup_i E|\alpha_i| < K$.*

Assumption 5 *The unobserved $m \times 1$ factor loadings, $\boldsymbol{\eta}_i$, for $i = 1, 2, ..., N$ are distributed independently of u_{jt}, and the common factor, $\mathbf{f}_t$, for all i, j and t, and are independently and identically distributed across i with zero means, and a finite covariance matrix, namely,*

$$\boldsymbol{\eta}_i \sim IID(\mathbf{0}, \boldsymbol{\Omega}_\eta), \tag{4}$$

where $\boldsymbol{\Omega}_\eta$ is an $m \times m$ symmetric positive definite matrix with $\|\boldsymbol{\Omega}_\eta\| < K$ and $\sup_i E\|\boldsymbol{\eta}_i\|^{4+\epsilon} < K$.

The above assumptions are standard in the literature on short T dynamic panels. Assumption 2 is innocuous and requires time effects and the factors to be time-varying. Note that the case where $\delta_t = \delta$ and/or $\mathbf{f}_t = \mathbf{f}$ for all t is already covered by the presence of the fixed-effects, α_i. Assumption 3 requires the regressors to be strictly exogenous with respect to ζ_{it}. This can be relaxed by considering a vector autoregressive version of (1) and (2) where $\mathbf{z}_{it} = (y_{it}, \mathbf{x}'_{it})'$ is modelled jointly as in Holtz-Eakin et al. (1988) and Binder et al. (2005). While in practice the choice of strictly exogenous variables is typically driven by economic theory and *prior* knowledge, tests for strict exogeneity are also available, see for example Su et al. (2016). Regarding possible correlation between $\boldsymbol{\eta}_i$ and the regressors $\Delta\mathbf{x}_i$, this can be controlled for by using the methods of Mundlak (1978) and Chamberlain (1982). Furthermore, while the composite error term, ζ_{it}, in (1) is cross-sectionally heteroskedastic through the presence of the interactive effects, allowing explicitly for the same in the idiosyncratic error, u_{it}, of (2) can be pursued along the lines of Hayakawa and Pesaran (2015). These authors extend the cross-sectionally independent homoskedastic idiosyncratic errors of Hsiao et al. (2002) to the heterosketastic case. The above extensions are not considered here as they are beyond the scope of the present focus of the paper. Assumption 4 permits a very general specification of fixed effects, which is one of the main strengths of the proposed method for empirical applications where little is known about the individual effects. Assumption 5 is required for identification of the factors and the parameters.

Combining (1) and (2), and eliminating the individual effects by first-differencing we have

$$\Delta y_{it} = \gamma \Delta y_{i,t-1} + \boldsymbol{\beta}' \Delta \mathbf{x}_{it} + d_t + \mathbf{g}_t' \boldsymbol{\eta}_i + \Delta u_{it}, \text{ for } t = 2, 3, ..., T;\ i = 1, 2, ..., N, \tag{5}$$

where $d_t = \Delta \delta_t \neq 0$ and $\mathbf{g}_t = \Delta \mathbf{f}_t \neq \mathbf{0}$ for some $t \geq 2$, and

$$\xi_{it} = \mathbf{g}_t' \boldsymbol{\eta}_i + \Delta u_{it}, \text{ for } t = 2, 3, ..., T. \tag{6}$$

For the specification of Δy_{i1} we make the following assumption about the initialization of (5):

Assumption 6 *Suppose that for each i, $\{\Delta y_{it}\}$ is started from time $t = -S + 1$, for some $S > 0$, with the initial first differences, $\Delta y_{i,-S+1}$, as random draws from a distribution such that*

$$E\left(\Delta y_{i,-S+1} \left| \Delta \mathbf{x}_i \right.\right) = a_S + \boldsymbol{\pi}_S' \Delta \mathbf{x}_i, \tag{7}$$

where $\Delta \mathbf{x}_i = (\Delta \mathbf{x}_{i1}', \Delta \mathbf{x}_{i2}', ..., \Delta \mathbf{x}_{iT}')'$ is the $kT \times 1$ vector of observations on the regressors, a_S is a fixed coefficient that allows for non-zero means, and $\boldsymbol{\pi}_S$ is the $kT \times 1$ vector of coefficients, such that $\sup_S |a_S| < K$, and $\sup_S \|\boldsymbol{\pi}_S\| < K$. Furthermore, let

$$\varpi_i = \Delta y_{i,-S+1} - E\left(\Delta y_{i,-S+1} \left| \Delta \mathbf{x}_i \right.\right), \tag{8}$$

and suppose that $\varpi_i \sim IID(0, \sigma_\varpi^2)$, $0 < \sigma_\varpi^2 < K$, and $\sup_i E\left|\varpi_i\right|^{4+\epsilon} < K$.

This assumption is not that restrictive and allows the initial values, $y_{i,-S}$ and $y_{i,-S+1}$ to depend on the fixed effects, α_i. Also it is redundant if $|\gamma| < 1$ and S is sufficiently large, and obviously does not apply if there are no regressors in (1). The main restriction here is the assumed linearity of (7).

The following proposition summarises the result for Δy_{i1}.

Proposition 1 *Under Assumptions 1, 3 and 6*

$$\Delta y_{i1} = d_1 + \boldsymbol{\pi}' \Delta \mathbf{x}_i + \xi_{i1}, \text{ for } i = 1, 2, ..., N, \tag{9}$$

where d_1 and $\boldsymbol{\pi}$ are unknown parameters of dimensions 1 and kT, respectively, and ξ_{i1} is the composite error defined by

$$\xi_{i1} = \tilde{\mathbf{g}}_1' \boldsymbol{\eta}_i + v_{i1}, \tag{10}$$

where $\|\tilde{\mathbf{g}}_1\| < K$. *The component* v_{i1} *is distributed independently of* $\Delta \mathbf{x}_i$ *and* $\boldsymbol{\eta}_i$ *and satisfies*

$$v_{i1} \sim IID(0, \omega\sigma^2), \quad \sup_i E\left|v_{i1}\right|^{4+\epsilon} < K, \tag{11}$$

for some small $\epsilon > 0$, *and a fixed* $K > 0$, *and*

$$Cov\left(v_{i1}, \Delta u_{it}\right) = \begin{cases} -\sigma^2 & \text{for } t = 2 \\ 0 & \text{for } t = 3, 4, ..., T \end{cases}, \tag{12}$$

where $0 < \omega_{\min} < \omega < \omega_{\max} < \infty$, *and* $\omega_{\min}$ *and* $\omega_{\max}$ *are fixed constants.*

Remark 1 *In the case where* $|\gamma| < 1$ *and* $S \to \infty$ *we have*

$$\Delta y_{i1} = d_1 + \boldsymbol{\pi}' \Delta \mathbf{x}_i + \xi_{i1},$$

where ξ_{i1}*is defined by (10), with* v_{i1} *given by*

$$v_{i1} = \sum_{j=0}^{\infty} \gamma^j \Delta u_{i,1-j} + \chi_i,$$

where

$$\chi_i = \sum_{j=0}^{\infty} \gamma^j \boldsymbol{\beta}' \Delta \mathbf{x}_{i,1-j} - E\left(\sum_{j=0}^{\infty} \gamma^j \boldsymbol{\beta}' \Delta \mathbf{x}_{i,1-j} \left|\Delta \mathbf{x}_i\right.\right).$$

Since $\Delta \mathbf{x}_{it}$, $\boldsymbol{\eta}_i$, *and* $u_{it'}$ *are independently distributed for all* i, t *and* t', *it then follows that* v_{i1} *is distributed independently of* $\boldsymbol{\eta}_i$ *and* $\Delta \mathbf{x}_i$, *with* $E\left(v_{i1}\right) = 0$, *and*

$$\begin{aligned} Var\left(v_{i1}\right) &= Var\left(\sum_{j=0}^{\infty} \gamma^j \Delta u_{i,1-j}\right) + Var\left(\chi_i\right) \\ &= \frac{2\sigma^2}{1+\gamma} + Var\left(\chi_i\right) > 0. \end{aligned}$$

In the case of pure AR(1) panels, we have the further parametric restriction, $Var\left(v_{i1}\right) = \frac{2\sigma^2}{1+\gamma}$, *which if imposed can increase estimation efficiency.*

Writing (5) and (9) in matrix notation we now have

$$\Delta \mathbf{y}_i = \Delta \mathbf{W}_i \boldsymbol{\varphi} + \boldsymbol{\xi}_i, \ \boldsymbol{\xi}_i = \mathbf{G}\boldsymbol{\eta}_i + \mathbf{r}_i, \tag{13}$$

where $\Delta \mathbf{y}_i = (\Delta y_{i1}, \Delta y_{i2}, ..., \Delta y_{iT})'$, $\Delta \mathbf{W}_i$ is the $T \times (Tk + 1 + k + T)$ matrix given by

$$\Delta \mathbf{W}_i = \begin{pmatrix} 1 & 0 & \dots & 0 & \Delta \mathbf{x}_i' & 0 & 0 \\ 0 & 1 & \dots & 0 & \mathbf{0} & \Delta \mathbf{x}_{i2}' & \Delta y_{i1} \\ \vdots & \vdots & \dots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \mathbf{0} & \Delta \mathbf{x}_{iT}' & \Delta y_{i,T-1} \end{pmatrix}, \tag{14}$$

$\boldsymbol{\varphi} = (\mathbf{d}', \boldsymbol{\pi}', \boldsymbol{\beta}', \gamma)'$ with $\mathbf{d} = (d_1, d_2, ..., d_T)'$, $\mathbf{G}' = (\tilde{\mathbf{g}}_1, \mathbf{g}_2, ..., \mathbf{g}_T)$, $\mathbf{r}_i = (v_{i1}, \Delta u_{i2}, ..., \Delta u_{iT})'$, and $\boldsymbol{\xi}_i = \left(\tilde{\xi}_{i1}, \xi_{i2}, \cdots, \xi_{iT}\right)'$, and recall that $\tilde{\xi}_{i1} = \tilde{\mathbf{g}}_1' \boldsymbol{\eta}_i + v_{i1}$, and $\xi_{it} = \mathbf{g}_t' \boldsymbol{\eta}_i + \Delta u_{it}$, for $t = 2, 3, ..., T$.

Proposition 2 *Consider the composite random variable, ξ_{it}, $i = 1, 2, ..., N$, for $t = 1$ defined by (10), and for $t = 2, 3, ..., T$ defined by (6). Then under Assumptions 1, 2, 3, 5, and 6, the following moment conditions hold:*

$$\sup_i E\left(|\xi_{it}|^{4+\epsilon}\right) < K, \text{ for } t = 1, 2, ..., T, \tag{15}$$

and

$$\sup_{i,t} E\left(\|\Delta \mathbf{x}_{it}\|^{4+\epsilon}\right) < K. \tag{16}$$

4 Quasi Maximum Likelihood Estimation

Consider the panel data model given by (13) and note that under Assumption 1, and using (11) and (12), we have

$$E(\mathbf{r}_i\mathbf{r}_i') = \sigma^2\mathbf{\Omega}, \tag{17}$$

where

$$E(\mathbf{r}_i\mathbf{r}_i') = \sigma^2 \begin{pmatrix} \omega & -1 & & & 0 \\ -1 & 2 & \ddots & & 0 \\ & & \ddots & & \\ & & \ddots & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix} = \sigma^2\mathbf{\Omega}, \tag{18}$$

and $\mathbf{\Omega} = \mathbf{\Omega}(\omega)$. Since $|\mathbf{\Omega}| = 1 + T(\omega - 1)$, ω needs to satisfy $\omega > 1 - \frac{1}{T}$ to ensure that $\mathbf{\Omega}$ is positive definite. Also, since $\boldsymbol{\eta}_i$ and $\mathbf{r}_i$ are independently distributed we have

$$Var(\boldsymbol{\xi}_i) = E(\boldsymbol{\xi}_i\boldsymbol{\xi}_i') = \sigma^2\mathbf{\Omega} + \mathbf{G}\mathbf{\Omega}_\eta\mathbf{G}' = \sigma^2(\mathbf{\Omega} + \mathbf{Q}\mathbf{Q}') = \mathbf{\Sigma}_\xi(\boldsymbol{\psi}) \tag{19}$$

where $\mathbf{Q} = (1/\sigma)\mathbf{G}\mathbf{\Omega}_\eta^{1/2}$, $rank(\mathbf{Q}) = m$, and $\boldsymbol{\psi} = (\omega, \sigma^2, vec(\mathbf{Q})')'$. With this normalisation, the quasi-log-likelihood of the transformed model (13) is given by

$$\begin{aligned} \ell_N(\boldsymbol{\theta}) &= \ell_N(\boldsymbol{\varphi}, \boldsymbol{\psi}) = -\frac{NT}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{\Sigma}_\xi(\boldsymbol{\psi})| - \frac{1}{2}\sum_{i=1}^N \boldsymbol{\xi}_i'(\boldsymbol{\varphi})\mathbf{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}) & (20) \\ &= -\frac{NT}{2}\ln(2\pi) - \frac{NT}{2}\ln(\sigma^2) - \frac{N}{2}\ln|\mathbf{\Omega} + \mathbf{Q}\mathbf{Q}'| - \frac{1}{2\sigma^2}\sum_{i=1}^N \boldsymbol{\xi}_i'(\boldsymbol{\varphi})(\mathbf{\Omega} + \mathbf{Q}\mathbf{Q}')^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}), & (21) \end{aligned}$$

where

$$\boldsymbol{\xi}_i(\boldsymbol{\varphi}) = \Delta\mathbf{y}_i - \Delta\mathbf{W}_i\boldsymbol{\varphi}, \tag{22}$$

and it is assumed that $\boldsymbol{\varphi}$ does not depend on $\boldsymbol{\psi}$. For fixed m and T, the above log-likelihood function depends on a fixed number of unknown parameters collected in the $[T(m+k+1)+k+3] \times 1$ vector $\boldsymbol{\theta} = (\boldsymbol{\varphi}', \boldsymbol{\psi}')'$.

To obtain the QML estimator, since $\mathbf{\Omega}$ is a positive definite matrix and $\mathbf{Q}\mathbf{Q}'$ is rank deficient (recall that by assumption $m < T$), we first note that

$$|\mathbf{\Omega} + \mathbf{Q}\mathbf{Q}'| = |\mathbf{\Omega}|\left|\mathbf{I}_m + \mathbf{Q}'\mathbf{\Omega}^{-1}\mathbf{Q}\right|,$$

and using the Woodbury matrix identity

$$\begin{aligned}(\mathbf{\Omega}+\mathbf{Q}\mathbf{Q}')^{-1} &= \mathbf{\Omega}^{-1}-\mathbf{\Omega}^{-1}\mathbf{Q}(\mathbf{I}_m+\mathbf{Q}'\mathbf{\Omega}^{-1}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{\Omega}^{-1} \\ &= \mathbf{\Omega}^{-1}-\mathbf{\Omega}^{-1}\mathbf{Q}\mathbf{A}^{-1}\mathbf{Q}'\mathbf{\Omega}^{-1},\end{aligned} \tag{23}$$

where $\mathbf{A}$ is a non-singular matrix defined by

$$\mathbf{A}=\mathbf{I}_m+\mathbf{Q}'\mathbf{\Omega}^{-1}\mathbf{Q}. \tag{24}$$

Using the above results in (21), and after some simplification the quasi-log-likelihood function can be written as

$$N^{-1}\ell_N(\boldsymbol{\theta}) \propto -\frac{T}{2}\ln(\sigma^2)-\frac{1}{2}\ln|\mathbf{\Omega}|-\frac{1}{2}\ln|\mathbf{A}|-\frac{1}{2\sigma^2}\left[Tr\left(\mathbf{B}_N\mathbf{\Omega}^{-1}\right)-Tr\left(\mathbf{B}_N\mathbf{\Omega}^{-1}\mathbf{Q}\mathbf{A}^{-1}\mathbf{Q}'\mathbf{\Omega}^{-1}\right)\right], \tag{25}$$

where $|\mathbf{\Omega}|=1+T(\omega-1)$, and

$$\mathbf{B}_N(\boldsymbol{\varphi})=N^{-1}\sum_{i=1}^{N}\boldsymbol{\xi}_i(\boldsymbol{\varphi})\boldsymbol{\xi}_i'(\boldsymbol{\varphi}). \tag{26}$$

If $\boldsymbol{\eta}_i$ and u_{it} are normally distributed, maximising (21) gives the maximum likelihood estimator of $\boldsymbol{\theta}$. If they are instead IID with mean zero and u_{it} has finite fourth moments, maximising (21) gives the QMLE of $\boldsymbol{\theta}$ (White 1982). Detailed regularity conditions can be found in Section 6.

For analytical convenience and identification purposes, which will become clearer below, we further define $\mathbf{P}=\mathbf{\Omega}^{-1/2}\mathbf{Q}\mathbf{A}^{-1/2}$. Note that since $\mathbf{A}$ and $\mathbf{\Omega}$ are non-singular matrices, then $rank(\mathbf{P})=m$, as well. Further, it is easily seen that

$$\mathbf{I}_m-\mathbf{P}'\mathbf{P}=\mathbf{I}_m-\mathbf{A}^{-1/2}\mathbf{Q}'\mathbf{\Omega}^{-1}\mathbf{Q}\mathbf{A}^{-1/2},$$

and using $\mathbf{Q}'\mathbf{\Omega}^{-1}\mathbf{Q}=\mathbf{A}-\mathbf{I}_m$ from (24), we have

$$\mathbf{A}^{-1}=\mathbf{I}_m-\mathbf{P}'\mathbf{P}. \tag{27}$$

Similarly,

$$Tr\left(\mathbf{B}_N\mathbf{\Omega}^{-1}\mathbf{Q}\mathbf{A}^{-1}\mathbf{Q}'\mathbf{\Omega}^{-1}\right)=\sigma^2 Tr\left[\mathbf{P}'\mathbf{C}_N(\boldsymbol{\phi})\mathbf{P}\right],$$

where

$$\mathbf{C}_N(\boldsymbol{\phi})=\sigma^{-2}\mathbf{\Omega}^{-1/2}\mathbf{B}_N(\boldsymbol{\varphi})\mathbf{\Omega}^{-1/2}, \tag{28}$$

and $\boldsymbol{\phi}=(\boldsymbol{\varphi}',\omega,\sigma^2)'$.

Using the above results, the quasi-log-likelihood function given by (25) can now be written as

$$N^{-1}\ell_N(\boldsymbol{\phi},\mathbf{P}) \propto -\frac{T}{2}\ln(\sigma^2)-\frac{1}{2}\ln\left[1+T(\omega-1)\right]+\frac{1}{2}\ln|\mathbf{I}_m-\mathbf{P}'\mathbf{P}|-\frac{1}{2}\left\{Tr\left[\mathbf{C}_N(\boldsymbol{\phi})\right]-Tr\left[\mathbf{P}'\mathbf{C}_N(\boldsymbol{\phi})\mathbf{P}\right]\right\}. \tag{29}$$

While as mentioned earlier the transformation from $\mathbf{Q}$ to $\mathbf{P}$ is carried out for analytical convenience, $\mathbf{P}$ is still not identified. It is easily seen that the value of $\ell_N(\boldsymbol{\phi},\mathbf{P})$ is invariant to the orthonormal transformation of $\mathbf{P}$. To see this consider the transformation $\tilde{\mathbf{P}}=\mathbf{P}\Xi$, where Ξ is an $m \times m$ orthonormal matrix such that $\Xi'\Xi = \mathbf{I}_m$. Then it is readily verified that $N^{-1}\ell_N(\boldsymbol{\phi}, \mathbf{P}) = N^{-1}\ell_N\left(\boldsymbol{\phi}, \tilde{\mathbf{P}}\right)$. Hence, $\mathbf{P}$ (or $\tilde{\mathbf{P}}$) is identified only up to an $m \times m$ orthonormal rotation matrix. Let $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_m)$, where $\mathbf{p}_t$ is the t^{th} column of $\mathbf{P}$, and $\mathbf{p}_t$ is a $T \times 1$ vector of unknown parameters. Since $rank(\mathbf{P}) = m$, then $\mathbf{P}'\mathbf{P}$ can be diagonalised by an orthonormal transformation, and without loss of generality we can impose the following $m(m-1)/2$ orthogonality conditions

$$\mathbf{p}_t'\mathbf{p}_s = 0, \text{ for all } s \neq t = 1, 2, ..., m. \tag{30}$$

Under these restrictions the quasi-log-likelihood function, (29), simplifies to

$$N^{-1}\ell_N(\boldsymbol{\phi}, \mathbf{P}) \propto -\frac{T}{2}\ln(\sigma^2) - \frac{1}{2}\ln\left[1 + T(\omega - 1)\right] + \frac{1}{2}\sum_{t=1}^{m}\ln(1 - \mathbf{p}_t'\mathbf{p}_t) + \frac{1}{2}\sum_{t=1}^{m}\mathbf{p}_t'\mathbf{C}_N(\boldsymbol{\phi})\mathbf{p}_t - \frac{1}{2}Tr\left[\mathbf{C}_N(\boldsymbol{\phi})\right]. \tag{31}$$

Taking first derivatives with respect to $\mathbf{p}_t$ and setting these derivatives to zero now yields

$$\mathbf{C}_N(\boldsymbol{\phi})\hat{\mathbf{p}}_t - \left(\frac{1}{1 - \hat{\mathbf{p}}_t'\hat{\mathbf{p}}_t}\right)\hat{\mathbf{p}}_t = \mathbf{0}, \quad \text{for } t = 1, 2, ..., m, \tag{32}$$

where $\hat{\mathbf{p}}_t$ is the quasi-maximum likelihood estimator of $\mathbf{p}_t$ (in terms of $\boldsymbol{\phi}$). Therefore, $\hat{\mathbf{p}}_t$ is the eigenvector of $\mathbf{C}_N(\boldsymbol{\phi})$ associated with the first m largest non-zero eigenvalues of $\mathbf{C}_N(\boldsymbol{\phi})$, which we denote by $\lambda_1(\boldsymbol{\phi}) > \lambda_2(\boldsymbol{\phi}) > .... > \lambda_m(\boldsymbol{\phi}) > 0$. Note that $\mathbf{C}_N(\boldsymbol{\phi})$ is a symmetric positive definite matrix with all real eigenvalues $\lambda_t(\boldsymbol{\phi}) > 0$, for $t = 1, 2, ..., T$. We also have

$$\lambda_t(\boldsymbol{\phi}) = \frac{1}{1 - \hat{\mathbf{p}}_t'\hat{\mathbf{p}}_t}, \quad \text{and} \quad \hat{\mathbf{p}}_t'\mathbf{C}_N(\boldsymbol{\phi})\hat{\mathbf{p}}_t = \lambda_t(\boldsymbol{\phi}) - 1.$$

Hence, the concentrated quasi-log-likelihood function in terms of $\boldsymbol{\phi}$ can be written as

$$N^{-1}\ell_N(\boldsymbol{\phi}; m) \propto -\frac{T}{2}\ln(\sigma^2) - \frac{1}{2}\ln\left[1 + T(\omega - 1)\right] - \frac{1}{2}\sum_{t=1}^{m}\ln\left[\lambda_t(\boldsymbol{\phi})\right] + \frac{1}{2}\sum_{t=1}^{m}\left[\lambda_t(\boldsymbol{\phi}) - 1\right] - \frac{1}{2}\sum_{t=1}^{T}\lambda_t(\boldsymbol{\phi}), \tag{33}$$

where $\lambda_t(\boldsymbol{\phi})$ is the t^{th} eigenvalue of $\mathbf{C}_N(\boldsymbol{\phi})$, given by (28). This concentrated quasi-log-likelihood function can now be maximised with respect to $\boldsymbol{\phi} = (\boldsymbol{\varphi}', \omega, \sigma^2)'$. The QML estimators, $\hat{\lambda}_t(\boldsymbol{\phi})$, can then be computed using the QML estimator of $\boldsymbol{\phi}$ and their corresponding variance covariance matrix can be computed using the delta method.

With regard to the computation of $\hat{\mathbf{p}}_t$ it is important to bear in mind that standard eigenvector routines provide eigenvectors that are typically orthonormalised. Whilst in the above analysis, $\hat{\mathbf{p}}_1$, $\hat{\mathbf{p}}_2, ..., \hat{\mathbf{p}}_m$ are orthogonal to each other, their length is not unity and is given by

$$\hat{\mathbf{p}}_t'\hat{\mathbf{p}}_t = 1 - \frac{1}{\lambda_t(\boldsymbol{\phi})}. \tag{34}$$

5 Identification conditions

We shall first derive necessary order conditions on m and T for identification, and then subject to these order conditions we derive additional conditions under which the parameters are locally identified, and show that global identification of short T panels with an error multi-factor structure is possible only in the case of panels with strictly exogenous regressors.

We begin our investigation by considering the order condition for identification of the panel AR(1) model. Using (5) and (9), we note that in this case

$$\begin{aligned}\Delta y_{it} &= d_t + \tilde{\mathbf{g}}_t'\boldsymbol{\eta}_i + v_{it}, \text{ for } t=1, \\ \Delta y_{it} - \gamma \Delta y_{i,t-1} &= d_t + \mathbf{g}_t'\boldsymbol{\eta}_i + \Delta u_{it}, \text{ for } t = 2,3,....,T,\end{aligned}$$

which can be written as

$$\mathbf{B}(\gamma)\Delta\mathbf{y}_i = \mathbf{d} + \mathbf{G}\boldsymbol{\eta}_i + \mathbf{r}_i = \mathbf{d} + \boldsymbol{\xi}_i, \text{ for } i = 1,2,...,N,$$

where $\mathbf{d} = (d_1, ..., d_T)'$, $\Delta\mathbf{y}_i$ and $\boldsymbol{\xi}_i$ are as defined above, and

$$\mathbf{B}(\gamma) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\gamma & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\gamma & 1 \end{pmatrix}. \tag{35}$$

Note also that, $|\mathbf{B}(\gamma)| = 1$, and

$$\mathbf{B}^{-1}(\gamma) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \gamma & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ \gamma^{T-1} & \cdots & \gamma & 1 \end{pmatrix}, \tag{36}$$

and hence

$$\Delta\mathbf{y}_i = \mathbf{a} + \mathbf{B}^{-1}(\gamma)\boldsymbol{\xi}_i,$$

where

$$\mathbf{a} = \mathbf{B}^{-1}(\gamma)\mathbf{d} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \gamma & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ \gamma^{T-1} & \cdots & \gamma & 1 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_T \end{pmatrix} = \begin{pmatrix} d_1 \\ \gamma d_1 + d_2 \\ \vdots \\ \gamma^{T-1}d_1 + \gamma^{T-2}d_2 + ... + \gamma d_{T-1} + d_T \end{pmatrix}. \tag{37}$$

Since $\mathbf{d}$ is a $T \times 1$ unrestricted parameter vector, then $\mathbf{a}$ is also unrestricted, namely knowing $\mathbf{a}$ does not help identify γ. Therefore, γ can only be identified from the $T(T+1)/2$ distinct elements of $Var(\Delta\mathbf{y}_i)$ which is given by

$$\begin{aligned}Var(\Delta\mathbf{y}_i) &= \mathbf{B}(\gamma)^{-1}Var(\boldsymbol{\xi}_i)\mathbf{B}'(\gamma)^{-1} \\ &= \sigma^2\mathbf{B}(\gamma)^{-1}(\boldsymbol{\Omega} + \mathbf{Q}\mathbf{Q}')\mathbf{B}'(\gamma)^{-1} = \boldsymbol{\Sigma}(\boldsymbol{\varrho}, \mathbf{Q}),\end{aligned}$$

where $\boldsymbol{\varrho} = (\gamma, \omega, \sigma^2)'$. But since $\mathbf{Q}$ enters $\boldsymbol{\Sigma}(\boldsymbol{\varrho}, \mathbf{Q})$ as $\mathbf{A} = \mathbf{Q}\mathbf{Q}'$ we need to consider the unknown elements of the symmetric matrix $\mathbf{A}$ under different rank conditions. First it is clear that if $\mathbf{A}$ has full rank, namely if $rank(\mathbf{A}) = T$, then $\boldsymbol{\varrho}$ is not identified. Hence, for identification of $\boldsymbol{\varrho}$, we must have $rank(\mathbf{A}) = rank(\mathbf{Q}) = m < T$. When $rank(\mathbf{Q}) = m$,

$\mathbf{Q}$ is identified only up to an $m \times m$ non-singular transformation. However, the number of non-redundant parameters of $\mathbf{Q}$ is given by $mT - m(m-1)/2$ (see p. 507 of Hayashi et al. (2007)). Hence, the order condition for identification of $\boldsymbol{\varrho}$ and the non-redundant elements of $\mathbf{Q}$ is given by

$$T(T+1)/2 \geq 3 + Tm - m(m-1)/2. \tag{38}$$

This order condition is satisfied if $T \geq 3$, for $m = 0, 1, 2, .., m_{\max}$ where $m_{\max}$ is the largest value of m that satisfies (38), that is $m_{\max} = T - 2$. It is easily seen that the above condition is not satisfied if $m = T - 1$. The maximized log-likelihood values for the rank deficient cases, $m = 0, 1, 2..., m_{\max}$ can be computed using (33).

Consider the more general case where the panel AR(1) model also contains exogenous regressors, and note that the system of equations (13) can be written equivalently as

$$\Delta \mathbf{y}_i = \mathbf{a} + \tilde{\mathbf{Z}}_i(\gamma)\,\boldsymbol{\delta} + \mathbf{B}^{-1}(\gamma)\,\boldsymbol{\xi}_i, \tag{39}$$

where $\mathbf{a}$, $\mathbf{B}^{-1}(\gamma)$ and $\boldsymbol{\xi}_i$ are as defined above, $\boldsymbol{\delta} = (\boldsymbol{\pi}', \boldsymbol{\beta}')'$, $\tilde{\mathbf{Z}}_i(\gamma) = \mathbf{B}^{-1}(\gamma)\,\mathbf{Z}_i$, and $\mathbf{Z}_i$ is the $T \times (Tk + k)$ matrix of observations on the exogenous regressors defined by

$$\mathbf{Z}_i = \begin{pmatrix} \Delta \mathbf{x}_i' & \mathbf{0} \\ \mathbf{0} & \Delta \mathbf{x}_{i2}' \\ \vdots & \vdots \\ \mathbf{0} & \Delta \mathbf{x}_{iT}' \end{pmatrix}. \tag{40}$$

It is clear from (39) that $\mathbf{a}$ and $\boldsymbol{\delta}$, and hence $\mathbf{d}$ and $\boldsymbol{\delta}$, are uniquely identified for a given value of γ. But it is already established that γ is identified from the covariance of $\mathbf{B}^{-1}(\gamma)\,\boldsymbol{\xi}_i$, given by $\boldsymbol{\Sigma}(\boldsymbol{\varrho}, \mathbf{Q}) = \sigma^2 \mathbf{B}(\gamma)^{-1}(\boldsymbol{\Omega} + \mathbf{Q}\mathbf{Q}')\,\mathbf{B}'(\gamma)^{-1}$, if the order condition (38) is met. Note that $\boldsymbol{\Sigma}(\boldsymbol{\varrho}, \mathbf{Q})$ does not depend on $\mathbf{d}$ and $\boldsymbol{\delta}$, and hence knowing $\mathbf{d}$ and $\boldsymbol{\delta}$ will not help identification of γ. As a result, the order condition (38) continues to be sufficient for identification of the parameters of the panel ARX(1) model.

To investigate necessary and sufficient conditions for identification of the parameters we consider the average log-likelihood function defined by (20) which we write as,

$$\bar{\ell}_N(\boldsymbol{\theta}) = N^{-1}\ell_N(\boldsymbol{\varphi}, \boldsymbol{\psi}) = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| - \frac{1}{2N}\sum_{i=1}^{N} \boldsymbol{\xi}_i'(\boldsymbol{\varphi})\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}), \tag{41}$$

where $\boldsymbol{\theta} = (\boldsymbol{\varphi}', \boldsymbol{\psi}')'$, $\boldsymbol{\varphi} = (\mathbf{d}', \boldsymbol{\pi}', \boldsymbol{\beta}', \gamma)' = (\boldsymbol{\lambda}', \gamma)'$, $\boldsymbol{\psi} = (\omega, \sigma^2, \mathbf{q}')'$, and $\mathbf{q}$ refers to the $[mT - m(m-1)/2] \times 1$ vector containing the non-reduntant elements of $\mathbf{Q}$. Suppose that $\boldsymbol{\lambda} \in \boldsymbol{\Theta}_\lambda$, $\gamma \in \boldsymbol{\Theta}_\gamma$, and $\boldsymbol{\psi} \in \boldsymbol{\Theta}_\psi$, and denote the true values of $\boldsymbol{\lambda}$, γ and $\boldsymbol{\psi}$ by $\boldsymbol{\lambda}_0, \gamma_0$, and $\boldsymbol{\psi}_0$, respectively. Consider the set $\mathcal{N}_\epsilon(\gamma_0)$ defined as follows:

Definition 1 *Let $\mathcal{N}_\epsilon(\gamma_0)$ be a set in the closed neighbourhood of γ_0 defined by*

$$\mathcal{N}_\epsilon(\gamma_0) = \{\gamma \in \boldsymbol{\Theta}_\gamma,\ |\gamma - \gamma_0| \leq \epsilon\},$$

for some small $\epsilon > 0$, where $\boldsymbol{\Theta}_\gamma$ is a compact subset of $\mathbb{R}$.

We now show that $\boldsymbol{\theta}_0 = (\boldsymbol{\varphi}_0', \boldsymbol{\psi}_0')' = (\boldsymbol{\lambda}_0', \gamma_0, \boldsymbol{\psi}_0')'$, where $\boldsymbol{\lambda} = (\mathbf{d}', \boldsymbol{\pi}', \boldsymbol{\beta}')'$ is identified on $\boldsymbol{\Theta}_\epsilon = \mathcal{N}_\epsilon(\gamma_0) \times \boldsymbol{\Theta}_\lambda \times \boldsymbol{\Theta}_\psi$. For this purpose, we require the following additional assumption.

Assumption 7 *(i)* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\epsilon} = \mathcal{N}_{\epsilon}(\gamma_0) \times \boldsymbol{\Theta}_{\lambda} \times \boldsymbol{\Theta}_{\psi}$*, where* $\boldsymbol{\Theta}_{\boldsymbol{\lambda}} = \boldsymbol{\Theta}_d \times \boldsymbol{\Theta}_{\pi} \times \boldsymbol{\Theta}_{\beta}$ *and* $\boldsymbol{\Theta}_{\psi} = \boldsymbol{\Theta}_{\omega} \times \boldsymbol{\Theta}_{\sigma} \times \boldsymbol{\Theta}_q$*,* $\boldsymbol{\Theta}_d$*,* $\boldsymbol{\Theta}_{\pi}$*,* $\boldsymbol{\Theta}_{\beta}$ *and* $\boldsymbol{\Theta}_q$ *are compact subsets of* $\mathbb{R}^{n_d}$*,* $\mathbb{R}^{n_{\pi}}$*,* $\mathbb{R}^{n_{\beta}}$*, and* $\mathbb{R}^{n_q}$*, respectively;* $\boldsymbol{\Theta}_{\omega}$ *and* $\boldsymbol{\Theta}_{\sigma}$ *are compact subsets of* $\mathbb{R}$*, where* $n_d = T$*,* $n_{\pi} = kT$*,* $n_{\beta} = k$*, and* $n_q = Tm - m(m-1)/2$*;* $\boldsymbol{\theta}_0 = (\boldsymbol{\varphi}_0', \boldsymbol{\psi}_0')' = (\boldsymbol{\lambda}_0', \gamma_0, \boldsymbol{\psi}_0')'$ *lies in the interior of* $\boldsymbol{\Theta}_{\epsilon}$ *(ii)* $\boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi}) = \sigma^2 (\boldsymbol{\Omega} + \mathbf{Q}\mathbf{Q}')$*, and for some* $c_{\max} > c_{\min} > 0$*,* $c_{\min} \leq \inf_{\boldsymbol{\psi} \in \boldsymbol{\Theta}_{\psi}} \lambda_{\min}[\boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi})] < \sup_{\boldsymbol{\psi} \in \boldsymbol{\Theta}_{\psi}} \lambda_{\max}[\boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi})] \leq c_{\max}$*, and (iii) as* $N \to \infty$

$$\mathbf{A}_N(\boldsymbol{\psi}) = \frac{1}{N} \sum_{i=1}^{N} \Delta \mathbf{W}_i' \boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi})^{-1} \Delta \mathbf{W}_i \overset{a.s.}{\to} \mathbf{A}(\boldsymbol{\psi}) \text{ uniformly in } \boldsymbol{\Theta}_{\psi}, \tag{42}$$

where $\mathbf{A}(\boldsymbol{\psi}) = \lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} E\left(\Delta \mathbf{W}_i' \boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi})^{-1} \Delta \mathbf{W}_i\right)$ *is positive definite for all values of* $\boldsymbol{\psi} \in \boldsymbol{\Theta}_{\psi}$*.*

The first part of this assumption is standard and rules out parameter values on the boundary of the parameter space, and since $\mathcal{N}_{\epsilon}(\gamma_0)$ is a subset of $\boldsymbol{\Theta}_{\gamma}$ which is compact, it also follows that $\boldsymbol{\Theta}_{\epsilon}$ being the Cartesian product of compact sets, is itself compact, namely $\boldsymbol{\Theta}_{\epsilon} \in \mathbb{R}^{n_{\theta}}$, where $n_{\theta} = 3 + T(k+1) + k + Tm - m(m-1)/2$. Note also that order condition (38) is taken into account in setting n_{θ}. The eigenvalue conditions on $\boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi})$ in the second part of the assumption are required for the proof of consistency results**.** This part of the assumption also holds when the order condition is met and $\omega > 1 - \frac{1}{T}$. Recall that under the latter $\boldsymbol{\Omega}$ is a positive definite matrix and $\mathbf{Q}$ is rank deficient, and under Assumption 1 $0 < \sigma^2 < K$. For γ we need to distinguish between the case where S is fixed (namely initialization is from a finite past) and when $S \to \infty$. Under the former it is only required that $|\gamma| < K$, which includes the unit root case ($|\gamma| = 1$). Under the latter (when $S \to \infty$), we must have $|\gamma| < 1$. Consider now the third part of Assumption 7, and note that

$$\sup_i E \left\| \Delta \mathbf{W}_i' \boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi})^{-1} \Delta \mathbf{W}_i \right\|^2 < \left\| \boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi})^{-1} \right\|^2 \sup_i E \left\| \Delta \mathbf{W}_i \right\|^4 < K,$$

where $\left\| \boldsymbol{\Sigma}_{\xi}(\boldsymbol{\psi})^{-1} \right\| < K$ under condition (ii) of Assumption 7, and $\sup_i E \left\| \Delta \mathbf{W}_i \right\|^4 < K$ by Lemma 1. It is also easily seen that $\Delta \mathbf{W}_i$ are cross-sectionally independent under Assumptions 1, 3, and 5. This follows since $\Delta \mathbf{x}_{it}$ are independent across i by Assumption 3, and Δy_{it} being a function of $\Delta \mathbf{x}_{it}$ and ξ_{it} (see (39)) are also cross-sectionally independent noting that ξ_{it} are cross-sectionally independent under 1 and 5. Hence, $\mathbf{A}_N(\boldsymbol{\psi}) \overset{a.s.}{\to} \mathbf{A}(\boldsymbol{\psi})$ for every $\boldsymbol{\psi} \in \boldsymbol{\Theta}_{\psi}$ (see, for example, Theorem 19.4 of Davidson (1994)). Under condition (ii) of Assumption 7 it is trivial to see that this result also holds uniformly in $\boldsymbol{\Theta}_{\psi}$. It is important to note that condition (42) holds even if we allow for common effects in $\mathbf{x}_{it}$ by relaxing Assumption 3 to allow for unobserved common factors so long as the factor loadings are cross-sectionally independent. Such effects are allowed for in the Monte Carlo experiments, see (52). Finally, the condition that $\mathbf{A}(\boldsymbol{\psi})$ is a positive definite matrix is needed for identification of $\boldsymbol{\varphi}$**.**

The main identification result is set out in the following proposition:

Proposition 3 *Consider the model given by (1) and (2), with the associated log-likelihood function for first-differences given by (20). Suppose that Assumptions 1-7, and the order condition (38) hold. Then* $\boldsymbol{\theta}_0$ *is almost surely locally identified for values of* γ *sufficiently close to* γ_0*, as formalised by definition 1.*

Remark 2 *In the absence of lagged dependent variables in (1), $\boldsymbol{\theta}_0$ is almost surely globally identified. This can be easily seen from the proof of Proposition 3 in the Appendix.*

6 Asymptotic properties of the QML estimator

The analysis of consistency and asymptotic normality of the QML estimator, $\hat{\boldsymbol{\theta}} = \arg\max_{\theta\in\boldsymbol{\Theta}_\epsilon} \bar{\ell}_N(\boldsymbol{\theta})$, now follows by application of standard results from the literature. Almost sure local consistency of $\hat{\boldsymbol{\theta}}$ follows, for example, from a straightforward adaptation of Theorem 9.3.1 of Davidson (2000). Specifically: (i) $\boldsymbol{\Theta}_\epsilon$ as a subset of $\boldsymbol{\Theta}$ is compact, (ii) setting $\bar{C}_N(\boldsymbol{\theta}) = -2\bar{\ell}_N(\boldsymbol{\theta})$, and $\bar{C}(\boldsymbol{\theta}) = E_0[\bar{C}_N(\boldsymbol{\theta})]$, under Assumptions 1-7, and using (A.39) and (A.40) we have that $\bar{C}_N(\boldsymbol{\theta}) \overset{a.s.}{\to} \bar{C}(\boldsymbol{\theta})$ uniformly on $\boldsymbol{\Theta}_\epsilon$ and (iii) $\boldsymbol{\theta}_0$ is the unique minimum of $\bar{C}(\boldsymbol{\theta})$ on $\boldsymbol{\Theta}_\epsilon$, and is an interior point of $\boldsymbol{\Theta}_\epsilon$, by assumption. Condition (iii) follows directly from condition (ii) and Proposition 3 (see Theorem 9.3.4., Davidson (2000)). Therefore, all three conditions of Theorem 9.3.1 of Davidson are satisfied and $\hat{\boldsymbol{\theta}} \overset{a.s.}{\to} \boldsymbol{\theta}_0$ on the set $\boldsymbol{\Theta}_\epsilon$.

The asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is derived by taking a Taylor expansion of $\frac{\partial\bar{\ell}_N(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}} = \mathbf{0}$ at $\boldsymbol{\theta}_0$ and checking the asymptotic behaviour of the score function, $\bar{\mathbf{s}}_N(\boldsymbol{\theta}) = \frac{\partial\bar{\ell}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$, and Hessian matrix, $\mathbf{H}_N(\boldsymbol{\theta}) = -\frac{\partial^2\bar{\ell}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$. If $E_0\left[\frac{\bar{\ell}_N(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right] = \mathbf{0}$ and $\mathbf{H}_N(\breve{\boldsymbol{\theta}}) \overset{a.s.}{\to} \mathbf{H}(\boldsymbol{\theta}_0)$ the asymptotic normality of the QMLE will follow from the mean value theorem:

$$\mathbf{0} = \sqrt{N}\bar{\mathbf{s}}_N(\hat{\boldsymbol{\theta}}) = \sqrt{N}\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0) - \mathbf{H}_N(\breve{\boldsymbol{\theta}})\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

where $\breve{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.

Let $\mathbf{J}_N(\boldsymbol{\theta}) = E_0\left[N\frac{\partial\bar{\ell}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\bar{\ell}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]$ be the variance-covariance matrix of the score vector. We state the following theorem.

Theorem 1 *Consider the dynamic panel data model given by (1) with interactive effects as in (2). Suppose that Assumptions 1 to 7, the order condition (38) and Proposition 3 hold. Denote the QML estimator of $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}} = \arg\max_{\theta\in\boldsymbol{\Theta}_\epsilon} \bar{\ell}_N(\boldsymbol{\theta})$, where $\bar{\ell}_N(\boldsymbol{\theta})$ is given by (41). Then, $\hat{\boldsymbol{\theta}}$ is almost surely locally consistent for $\boldsymbol{\theta}_0$ on $\boldsymbol{\Theta}_\epsilon$ for values of γ sufficiently close to γ_0 as formalised by definition 1, and*

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d N\left[\mathbf{0}, \mathbf{H}^{-1}(\boldsymbol{\theta}_0)\mathbf{J}(\boldsymbol{\theta}_0)\mathbf{H}^{-1}(\boldsymbol{\theta}_0)\right], \tag{43}$$

where $\mathbf{H}(\boldsymbol{\theta}_0) = \lim_{N\to\infty} E_0\left[-\frac{\partial^2\bar{\ell}_N(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]$ and $\mathbf{J}(\boldsymbol{\theta}_0) = \lim_{N\to\infty} E_0\left[N\frac{\partial\bar{\ell}_N(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\frac{\partial\bar{\ell}_N(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\right]$, both assumed to exist.

When $\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)$ is Gaussian $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d N\left[\mathbf{0}, \mathbf{H}^{-1}(\boldsymbol{\theta}_0)\right]$. A consistent estimator for the variance in (43) can be obtained by substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ in the expressions for $\mathbf{J}(\boldsymbol{\theta}_0)$ and $\mathbf{H}(\boldsymbol{\theta}_0)$.

7 Estimating the number of factors

There are a number of studies that provide information criteria for selecting the number of factors including Bai and Ng (2002), Onatski (2009), Kapetanios (2010), Ahn and Horenstein

(2013), among others. However, these are not applicable to short T panel data sets, and require both N and T to be large. In the case of short T panels Ahn et al. (2013) estimate the true number of factors, m_0, within a GMM framework using information criteria as well as the Sargan-Hansen misspecification statistic, in a sequential manner. To ensure consistency of the selected number of factors, following Bauer et al. (1988) and Cragg and Donald (1997), Ahn et, al. (2013) choose the significance level b_N such that $b_N \to 0$ and $-\ln(b_N)/N \to 0$ as $N \to \infty$. Using simulations they find that the sequential method could produce better estimates if the significance level depends also on T (in addition to N), when the regressors and individual effects are not highly correlated, but do not provide theoretical details on how best to allow for T as well as N in their selection procedure. In what follows we consider a sequential likelihood ratio (LR) testing procedure, but adjust the critical values of the tests to take account of the multiple testing nature of the procedure in terms of T, as well as adjusting the critical values of the tests in terms of N to ensure consistency of the selected number of factors. We provide a formal theory that should be of general interest for the analysis of short T factor models.

7.1 A sequential multiple testing likelihood ratio procedure for estimating the number of factors

Our sequential multiple testing likelihood ratio (MTLR) procedure makes use of the likelihood ratio statistic and in effect involves sequentially performing a number of likelihood ratio tests of the overidentifying restrictions on the model defined by (13). To see this, from (38) it follows that the degree of freedom (DF) for the test is given by

$$DF = T(T+1)/2 - (3 + Tm - m(m-1)/2), \tag{44}$$

and depends on m and T. When $m = m_{\max} = T-2$, $DF = 0$ and therefore the panel data model is exactly identified, and there are no free parameters (restrictions) to test. The LR tests involving over-identifying restrictions are defined by tests of $m = \{0, 1, 2, .., T-3\}$ against $m_{\max} = T-2$. Let $\hat{\boldsymbol{\theta}}_m$ be the QML estimator of $\boldsymbol{\theta}$, assuming m unobserved common factors, using the concentrated log-likelihood function given by (33) in terms of m and $\boldsymbol{\phi} = (\boldsymbol{\varphi}', \omega, \sigma^2)'$, which we reproduce here for convenience, making the dependence of $\boldsymbol{\phi}$ on m explicit:

$$\begin{aligned}\ell_N(\boldsymbol{\phi}_m; m) \;\propto\; & -\frac{TN}{2}\ln(\sigma_m^2) - \frac{N}{2}\ln\left[1 + T(\omega_m - 1)\right] - \frac{N}{2}\sum_{t=1}^{m}\ln\left[\lambda_t(\boldsymbol{\phi}_m)\right] \\ & + \frac{N}{2}\sum_{j=1}^{m}\left[\lambda_t(\boldsymbol{\phi}_m) - 1\right] - \frac{N}{2}\sum_{t=1}^{T}\lambda_t(\boldsymbol{\phi}_m),\end{aligned}$$

where $\lambda_1(\boldsymbol{\phi}_m) > \lambda_2(\boldsymbol{\phi}_m) > .... > \lambda_T(\boldsymbol{\phi}_m) > 0$ are the eigenvalues of $\mathbf{C}_N(\boldsymbol{\phi}_m) = \sigma^{-2}\boldsymbol{\Omega}_m^{-1/2}\mathbf{B}_N(\boldsymbol{\varphi}_m)\boldsymbol{\Omega}_m^{-1/2}$.[2] Then the LR statistics for testing H_0: $m = m_0$ against H_1: $m = m_{\max}$, for $m_0 = \{0, 1, 2, .., T-3\}$ and $m_{\max} = T-2 > m_0$, are given by

$$\mathcal{LR}_N(m_{\max}, m_0) = 2\left[\ell_N\left(\hat{\boldsymbol{\phi}}_{m_{\max}}; m_{\max}\right) - \ell_N\left(\hat{\boldsymbol{\phi}}_{m_0}; m_0\right)\right], \tag{45}$$

[2] Recall that $\mathbf{B}_N(\boldsymbol{\varphi}_m)$ is defined by (26), and hence $\mathbf{C}_N(\boldsymbol{\phi}_m)$ is a positive definite matrix.

where $\hat{\boldsymbol{\phi}}_m = \arg\max_{\boldsymbol{\phi}_m} \ell_N(\boldsymbol{\phi}_m; m)$. Under the assumption that $\boldsymbol{\xi}_i$ in (13) is Gaussian, and the panel data model is correctly specified with $m = m_0$, then using standard asymptotic results we have $\mathcal{LR}_N(m_{\max}, m_0) \to_d \chi^2_{DF}$, as $N \to \infty$ for a fixed T, where DF is given by (44) for the relevant choices of $m = m_{\max}$ and m_0. The following sequential testing procedure can now be adopted to estimate m:

$$\hat{m} = 0, \text{ if a test based on } \mathcal{LR}_N(m_{\max} = T-2, m_0 = 0) \text{ is not rejected.}$$

$$\begin{aligned} \hat{m} &= 1, \text{ if a test based on } \mathcal{LR}_N(m_{\max} = T-2, m_0 = 0) \text{ is rejected,} \\ &\quad \text{AND a test based on } \mathcal{LR}_N(m_{\max} = T-2, m_0 = 1) \text{ is not rejected.} \end{aligned}$$

$$\begin{aligned} \hat{m} &= 2, \text{ if a test based on } \mathcal{LR}_N(m_{\max} = T-2, m_0 = 0) \text{ and } \mathcal{LR}_N(m_{\max} = T-2, m_0 = 1) \\ &\quad \text{are both rejected AND a test based on } \mathcal{LR}_N(m_{\max} = T-2, m_0 = 2) \text{ is not rejected.} \end{aligned}$$

This sequential procedure is continued until $m_0 = T - 3$. Since $T - 2$ separate tests are carried out, to control the overall size of the sequential testing procedure we need to adjust the size of the underlying individual tests. As the true number of factors, m_0, is unknown and could be $T - 2$, in what follows we assume the sequential procedure involves $T - 2$ separate tests, although in some applications we might end up stopping the sequential procedure having carried out a fewer number of tests than $T - 2$. Let the null hypotheses of interest be $H_{T-2,0}, H_{T-2,1}, ..., H_{T-2,T-3}$, and write the $T - 2$ LR tests as

$$\Pr\left(\mathcal{LR}_N(m_{\max} = T-2, m_0 = t-1) > CV_{N,T-2,t-1} \left| H_{T-2,t-1} \right.\right) \leq p_{N,T-2,t-1}, \text{ for } t = 1, 2, ..., T-2,$$

where $CV_{N,T-2,t-1}$ is the critical value for the test of $H_{T-2,t-1}$, and $p_{N,T-2,t-1}$ is the realized p-value for $H_{T-2,t-1}$. The overall size of the test is now given by the family-wise error rate (FWER) defined by

$$FWER_N = \Pr\left[\cup_{t=1}^{T-2}\left(\mathcal{LR}_N(m_{\max} = T-2, m_0 = t-1) > CV_{N,T-2,t-1} \left| H_{T-2,t-1} \right.\right)\right].$$

Suppose that we wish to control $FWER_N$ to lie below a pre-determined value, α. An exact solution to this problem depends on the nature of the dependence across the underlying tests, which is generally difficult to obtain. But one could derive bounds on $FWER_N$ using, for example, the Bonferroni (1936) or Holm (1979) procedures. Both of these procedures are valid for all possible degrees of dependence across the individual tests, and as a result tend to be conservative in the sense that the actual size will be lower than the overall target size of α. Using Boole's inequality (also known as the union bound) we have

$$\begin{aligned} &\Pr\left\{\cup_{t=1}^{T-2}\left[\mathcal{LR}_N(m_{\max} = T-2, m_0 = t-1) > CV_{N,T-2,t-1} \left| H_{T-2,t-1} \right.\right]\right\} \\ \leq &\sum_{t=1}^{T-2} \Pr\left(\mathcal{LR}_N(m_{\max} = T-2, m_0 = t-1) > CV_{N,T-2,t-1} \left| H_{T-2,t-1} \right.\right) \\ \leq &\sum_{t=1}^{T-2} p_{N,T-2,t-1}. \end{aligned}$$

Hence, to obtain $FWER_N \leq \alpha$, it is sufficient to set $p_{N,T-2,t-1} \leq \alpha/(T-2)$. The individual critical values, $CV_{N,T-2,t-1}$ are based on the asymptotic critical values (as $N \to \infty$) of the

χ^2 distribution, namely $\chi^2_{DF}\left[\alpha/\left(T-2\right)\right]$, where $\alpha/(T-2)$ is the right-tail probability of the individual tests.

The above sequential MTLR procedure ensures that $\lim_{N\to\infty} FWER_N \leq \alpha$, but this by itself does not guarantee that m_0, the true value of m, will be estimated consistently. This is a well known problem in the sequential testing literature. To achieve consistency we need to allow α to decline with N at a suitable rate as will be shown in what follows.

Under non-normal errors the LR statistic defined by (45) need not be chi-squared distributed. This follows from known results for the likelihood ratio statistic under misspecification. See, for example, Foutz and Srivastava (1977) who show that under misspecification the LR statistic behaves asymptotically as a linear combination of independent chi-squared variates. This is also in line with results in Satorra and Bentler (1994) and Yuan and Bentler (2007) for standard factor models. Following this literature we conjecture that under non-Gaussian errors the null distribution of $\mathcal{LR}_N\left(m_{\max}, m_0\right)$ can also be asymptotically approximated as a linear combination of independent chi-squared variates. Simulation results reported in the online supplement confirm that $\mathcal{LR}_N\left(m_{\max}, m_0\right)$ is oversized when using chi-square critical values in this case. However, even under non-normal errors, the above sequential procedure using critical values of the chi-square distribution can still consistently estimate the true number of factors as shown in the following proposition and associated theorem.

Proposition 4 *Suppose under the null hypothesis H_0 the LR test statistic $\mathcal{LR}_N$ is distributed as $\sum_{i=1}^{k} w_i\chi_i^2(1)$, where the weights $w_1 \geq w_2 \geq ... \geq w_k > 0$ are finite constants, and $\chi_i^2(1)$ for $i = 1, 2, ..., k$ are independently distributed central chi-squared variates with 1 degree of freedom. Further suppose that under the alternative hypothesis H_1 $\mathcal{LR}_N$ is distributed as $\sum_{i=1}^{k} w_i\chi_i^2(1, \mu_{i,N}^2)$, where $\chi_i^2(1, \mu_{i,N}^2)$ for $i = 1, 2, ...,\ k$ are independently distributed non-central chi-squared variates with 1 degree of freedom and non-centrality parameter, $\mu_{i,N}^2$, $i = 1, 2, ..., k$. Denote the non-centrality parameter of the test under H_1 by $\mu_N^2 = \sum_{i=1}^{k}\mu_{i,N}^2$. Suppose k is a finite integer, and $\mu_N^2 = O(N)$. Denote type I and II errors of the test by α_N and β_N, respectively, and the critical value of the test by $c_N^2(k)$. Under Assumptions 1-7 if $c_N^2\left(k\right) \to \infty$ and $\mu_N^2 \to \infty$ as $N \to \infty$ such that $c_N^2\left(k\right)/\mu_N^2 \to 0$, then both α_N and $\beta_N \to 0$.*

Remark 3 *The standard chi-squared test is included in the above proposition as a special case by setting $w_i = 1$, for all i.*

Remark 4 *Clearly, the conditions of Proposition 4 are met if $\alpha_N = p/f(N)$, where $f(N) = N^{\delta}$, with δ a finite non-zero constant. Further, using (A.47) from the proof of Proposition 1 in the Appendix we have*

$$\frac{c_N^2(k)}{\mu_N^2} \leq \frac{2\theta_{\min}^{-2}\ln\left(\frac{k}{\alpha_N}\right)}{\mu_N^2} = \frac{2w_1 k\ln\left(\frac{kN^{\delta}}{p}\right)}{\mu_N^2} = O\left(\frac{\delta\ln(N)}{\mu_N^2}\right), \tag{46}$$

and since by assumption $\mu_N^2 = O(N)$ it follows that $c_N^2(k)/\mu_N^2 \to 0$ as required.

Remark 5 *When α_N is set as $\alpha_N = p/N^{\delta}$, the parameter p $(0 < p < 1)$ can be viewed as the nominal size of the test. Then $\beta_N \to 0$ if $\ln N/\mu_N^2 \to 0$, which is satisfied in the standard case where $\mu_N^2 = O(N)$. The Neyman-Pearson case is obtained if we set $\delta = 0$. The case of $\delta > 0$ relates to the Chernoff test procedure that aims at minimizing $\Pr(H_0)\alpha_N + \Pr(H_1)\beta_N$,* *where* $0 < \Pr(H_0) < 1$ *and* $0 < \Pr(H_1) < 1$ *are prior probabilities of* H_0 *and* H_1*, respectively. When* N *is finite the solution to this problem depends on the prior probabilities. But in the case of chi-squared tests, we have* $\Pr(H_0)\alpha_N + \Pr(H_1)\beta_N \to 0$ *as* $N \to \infty$*, irrespective of the prior probabilities* $\Pr(H_0)$ *and* $\Pr(H_1)$*, so long as* $\alpha_N = p/N^{\delta}$ *for* $\delta > 0$ *and* $p > 0$*.*

Remark 6 *In finite samples the choice of* p *and* δ *can matter, though for moderate values of* N *the choice of* p *is likely to be of second order importance. In the simulation results that follow we set* $\delta = 1$ *and* $p = 5\%$*.*

Theorem 2 *Let* $\hat{m}$ *be the number of factors obtained using the sequential likelihood ratio procedure based on the statistic* $\mathcal{LR}_N(m_{\max}, m_0)$ *given by (45) for which Proposition 4 holds. Then* $\Pr(\hat{m} = m_0) \to 1$*.*

From Proposition 4 and Theorem 2 it follows that $\hat{m}$ obtained using the sequential MTLR procedure described above is a consistent estimator of the true number of factors m_0. In line with the above discussion in the ensuing Monte Carlo results when performing the sequential MTLR procedure we use $\alpha_N = \frac{\kappa\, p}{(T-2)N^{\delta}}$, where κ is some positive constant such that condition (46) holds approximately.

8 Monte Carlo design and results

In this section, we investigate the finite sample properties of the proposed estimator using Monte Carlo (MC) simulations. We begin by presenting the MC designs that we shall be employing for the pure AR(1) panels and dynamic panels with regressors.

8.1 Monte Carlo design

8.1.1 The AR(1) model

The observations on y_{it} are generated assuming m unobserved factors as

$$\begin{aligned} y_{it} &= \alpha_i + \mu_\delta \delta_t + \gamma y_{i,t-1} + \zeta_{it}, \quad \text{for } i = 1,2,...,N; t = -S+1, -S+2, .., 0, 1, ..., T, \\ \zeta_{it} &= \boldsymbol{\eta}_i' \mathbf{f}_t + u_{it}, \end{aligned}$$

with the idiosyncratic errors generated as $u_{it} \sim IID\mathcal{N}(0, \sigma^2)$ under Gaussian errors, and $u_{it} \sim IID\frac{\sigma}{\sqrt{12}}(\chi^2_6 - 6)$ under non-Gaussian errors, where χ^2_6 is a chi-square variate with six degrees of freedom. In the case where $|\gamma| < 1$, we start the process with

$$y_{i,-S+1} = \frac{\alpha_i}{1-\gamma} + \sum_{j=0}^{S-1} \gamma^j \zeta_{i,-j},$$

and set $S = 50$ to reduce the impact of the initial values on the sample period used in the analysis, which we take to be $t = 0, 1, ..., T$. In the unit root case we initially generate the first-differences and then cumulate them to obtain y_{it} starting from some arbitrary values for y_{i0}, $i = 1, 2, ..., N$. The first-differences are generated as

$$\begin{aligned} \Delta y_{i1} &= \mu_\delta \Delta\delta_1 + \Delta\zeta_{i1} \\ \Delta y_{it} &= \mu_\delta \Delta\delta_t + \gamma \Delta y_{i,t-1} + \Delta\zeta_{it},\ t = 2, 3, ..., T, \end{aligned}$$

where the process is initalised at $\Delta y_{i0} = 0$, with $\gamma = 1$.[3]

For both the stationary and unit root cases, after first-differencing we end up with T observations that are used in estimation. The factor loadings are generated as

$$\eta_{\ell i} \sim IID\mathcal{N}(0, \sigma^2_{\eta\ell}),\ \ \ell = 1, 2, ..., m, \tag{47}$$

and the unobserved common factors, $f_{\ell t}$, as

$$f_{\ell t} = \rho_{f\ell} f_{\ell,t-1} + \sqrt{1 - \rho^2_{f\ell}}\varepsilon_{f\ell t},\ \varepsilon_{f\ell t} \sim IID\mathcal{N}(0, 1), \text{ for } \ell = 1, 2, ..., m; t = -S+1, ..., -1, 0, 1, .., T, \tag{48}$$

with $\rho_{f\ell} = 0.9$, and without loss of generality we set $f_{\ell,-S} = 0$ with $S = 50$ throughout. The resultant $f_{\ell t}$ values are re-scaled such that $T^{-1}\sum_{t=1}^{T} f^2_{\ell t} = 1$, for all ℓ. Specifically we impose the following normalisations on the common factors

$$T^{-1}\sum_{t=1}^{T} f_{\ell t} = 0,\ T^{-1}\sum_{t=1}^{T} f^2_{\ell t} = 1, \text{ and } T^{-1}\sum_{t=1}^{T} f_{\ell' t} f_{\ell t} = 0\,, \text{ for } \ell\ \neq \ell'. \tag{49}$$

We generate the time effects as $\delta_t = \frac{1}{2}(t^2 - t)$ which are further normalised so that

$$T^{-1}\sum_{t=1}^{T} \delta_t = 0,\ T^{-1}\sum_{t=1}^{T} \delta^2_t = 1, \text{ and } T^{-1}\sum_{t=1}^{T} \delta_t f_{\ell t} = 0, \text{ for all } \ell. \tag{50}$$

The fixed effects, α_i, are generated as

$$\alpha_i = b_1 \bar{u}_i + b_2 v_i,$$

where $\bar{u}_i = T^{-1}\sum_{t=1}^{T} u_{it}$, and $v_i \sim IID\mathcal{N}(0, 1)$. b_1 and b_2 are fixed constants to be set later. This set up ensures that the fixed effects are correlated with the idiosyncratic errors when $b_1 \neq 0$. The values of the remaining parameters are set as

$$\mu_\delta = 2\sigma,\ \sigma^2_{\eta\ell} = \sigma^2/m,\ \text{ for all } \ell.$$

Finally, as shown in Section A.3 of the Appendix, the average fit of the panel AR(1) model is determined by γ and does not depend on $\sigma^2 = Var(u_{it})$, and hence we set $\sigma^2 = 1$. For the key parameter of the model, γ, we consider a medium and a high value, namely $\gamma = 0.4$ and 0.8, as well as $\gamma = 1$.

We report simulation results for the autoregressive parameter γ for the following combinations of sample sizes, $T = \{5, 10\}$ and $N = \{100, 300, 500\}$. Specifically, we report the bias and root mean square error (RMSE). In addition, we present size and power estimates. Power is presented for $\gamma = \{0.30, 0.70, 0.96\}$ for the null values of $\gamma = \{0.4, 0.8, 1.0\}$. All tests are carried out at the 5% significance level and all experiments are replicated 2,000 times, unless otherwise stated.

8.1.2 The ARX(1) model

The observations on y_{it} for the panel ARX(1) model are generated assuming $k = 1$ (one exogenous regressor) and m unobserved factors as

$$\begin{aligned} y_{it} &= \alpha_i + \mu_\delta \delta_t + \gamma y_{i,t-1} + \beta x_{it} + \zeta_{it}, \quad \text{for } i = 1, 2, ..., N; t = -S+1, -S+2, .., 0, 1, ..., T, \\ \zeta_{it} &= \boldsymbol{\eta}_i' \mathbf{f}_t + u_{it}, \end{aligned} \tag{51}$$

[3] Any value for Δy_{i0} could be used and the results would not be affected since the value of Δy_{i0} gets absorbed in the intercept term of the underlying DGP.

with the idiosyncratic errors, the time effects and factors and associated loadings generated as for the AR(1) model. The values of σ^2 and $\sigma^2_{\eta\ell}$ are set below to ensure a certain degree of average fit for the panel regression in (51). For the case $|\gamma| < 1$, we initialize the DGP with

$$y_{i,-S+1} = \frac{\alpha_i}{1-\gamma} + \beta \sum_{j=0}^{S-1} \gamma^j x_{i,-j} + \sum_{j=0}^{S-1} \gamma^j \zeta_{i,-j},$$

where as before we set $S = 50$ and discard the first 49 observations. As for the AR(1) model, in the unit root case we begin with generating the first-differences and then cumulate them to obtain y_{it} from some arbitrary values for y_{i0}. The first-differences are generated as

$$\begin{aligned} \Delta y_{i1} &= \mu_\delta \Delta\delta_1 + \beta\Delta x_{i1} + \Delta\zeta_{i1} \\ \Delta y_{it} &= \mu_\delta \Delta\delta_t + \gamma\Delta y_{i,t-1} + \beta\Delta x_{it} + \Delta\zeta_{it},\ t = 2, 3, ..., T, \end{aligned}$$

with $\Delta y_{i0} = 0$ and $\gamma = 1$. In both cases the observations $t = 0$ through T are used for estimation, thus ending up with T observations for estimation after first-differencing.

The regressor, x_{it}, is generated as

$$x_{it} = \mu_i + \boldsymbol{\vartheta}_i'\mathbf{f}_t + \breve{x}_{it},\ ,\breve{x}_{it} = \rho_x \breve{x}_{i,t-1} + \sqrt{1-\rho_x^2}\varepsilon_{it}, \tag{52}$$

for $t = -S+1, ..., 0, 1, ..., T$, with $\breve{x}_{i,-S} = 0$, $|\rho_x| < 1$, $\mu_i \sim IID\mathcal{N}(0,1)$, $\boldsymbol{\vartheta}_i = (\vartheta_{1i}, \vartheta_{2i}, ..., \vartheta_{mi})'$, and $\varepsilon_{it} \sim IID\mathcal{N}(0,1)$. We set $\rho_x = 0.8$. The factor loadings, $\boldsymbol{\vartheta}_i$, in the x_{it} process are generated as

$$\vartheta_{\ell i} \sim IID\mathcal{N}(\mu_{\vartheta\ell}, \sigma^2_{\vartheta\ell}),\ \text{for } \ell = 1, 2, ..., m. \tag{53}$$

The fixed effects, α_i, are generated as

$$\alpha_i = b_0\bar{x}_i + b_1\bar{u}_i + b_2 v_i,$$

where $\bar{x}_i = T^{-1}\sum_{t=1}^T x_{it}$, $\bar{u}_i = T^{-1}\sum_{t=1}^T u_{it}$, $v_i \sim IID\mathcal{N}(0,1)$ and b_0, b_1, b_2 are fixed constants to be defined later. This set up ensures that the fixed effects are correlated both with the regressors and the idiosyncratic errors when $b_0 \neq 0$ and $b_1 \neq 0$.

We calibrate the rest of the parameters to ensure a given average measure of fit, as defined by the average R^2 derived in Section A.3 of the Appendix. In this way we ensure that the fit of the underlying model does not change with m, the number of factors. Using (51) and (52), for the case where $|\gamma| < 1$ we have

$$R_y^2 = \frac{\beta^2 Var(\breve{x}_{it}) + \gamma^2\left[\mu_\delta^2 + \sum_{\ell=1}^m \left(N^{-1}\sum_{i=1}^N c_{\ell i}^2\right) + \sigma^2\right]}{\beta^2 Var(\breve{x}_{it}) + \mu_\delta^2 + \sum_{\ell=1}^m \left(N^{-1}\sum_{i=1}^N c_{\ell i}^2\right) + \sigma^2},$$

where $c_{\ell i} = \beta\vartheta_{\ell i} + \eta_{\ell i}$. Also, in view of (52) we have $Var(\breve{x}_{it}) = 1$. Hence

$$R_y^2 = \frac{\beta^2 + \left(\mu_\delta^2 + N^{-1}\sum_{i=1}^N c_i^2 + \sigma^2\right)\gamma^2}{\beta^2 + \mu_\delta^2 + N^{-1}\sum_{i=1}^N c_i^2 + \sigma^2},$$

where $c_i^2 = \sum_{\ell=1}^m c_{\ell i}^2$. But

$$\begin{aligned} N^{-1}\sum_{i=1}^N c_i^2 &= \sum_{\ell=1}^m \left(N^{-1}\sum_{i=1}^N c_{\ell i}^2\right) = \sum_{\ell=1}^m N^{-1}\sum_{i=1}^N \left(\beta\vartheta_{\ell i} + \eta_{\ell i}\right)^2 \\ &= \beta^2 \sum_{\ell=1}^m \left(N^{-1}\sum_{i=1}^N \vartheta_{\ell i}^2\right) + \sum_{\ell=1}^m \left(N^{-1}\sum_{i=1}^N \eta_{\ell i}^2\right) \\ &\quad +2\beta\sum_{\ell=1}^m \left(N^{-1}\sum_{i=1}^N \eta_{\ell i}\vartheta_{\ell i}\right), \end{aligned}$$

and for N sufficiently large and noting that $\vartheta_{\ell i}$ and $\eta_{\ell i}$ are generated independently, we have (see (47) and (53))

$$N^{-1}\sum_{i=1}^N \vartheta_{\ell i}^2 \rightarrow_p Var\left(\vartheta_{\ell i}\right) + \left[E\left(\vartheta_{\ell i}\right)\right]^2 = \sigma_{\vartheta\ell}^2 + \mu_{\vartheta\ell}^2,$$

$$N^{-1}\sum_{i=1}^N \eta_{\ell i}^2 \rightarrow_p Var\left(\eta_{\ell i}\right) + \left[E\left(\eta_{\ell i}\right)\right]^2 = \sigma_{\eta\ell}^2,$$

and since $E\left(\eta_{\ell i}\right) = 0$, we also have $N^{-1}\sum_{i=1}^N \eta_{\ell i}\vartheta_{\ell i} \rightarrow_p 0$. Hence

$$N^{-1}\sum_{i=1}^N c_i^2 \rightarrow_p \beta^2 \sum_{\ell=1}^m \left(\sigma_{\vartheta\ell}^2 + \mu_{\vartheta\ell}^2\right) + \sum_{\ell=1}^m \sigma_{\eta\ell}^2.$$

Using the above results and setting $\beta = 1$ we obtain

$$R_y^2 - \gamma^2 = \frac{1-\gamma^2}{1 + \mu_\delta^2 + \sum_{\ell=1}^m \left(\sigma_{\vartheta\ell}^2 + \mu_{\vartheta\ell}^2\right) + \sum_{\ell=1}^m \sigma_{\eta\ell}^2 + \sigma^2}.$$

We control the value of $R_y^2 - \gamma^2$ to be the same for all values of m. To this end, the value of the remaining parameters are set as

$$\mu_\delta = 2\sigma,\ \sigma_{\eta\ell}^2 = \sigma_{\vartheta\ell}^2 = \sigma^2/m,\ \ \mu_{\vartheta\ell} = \sigma/\sqrt{m}, \text{ for all } \ell, \tag{54}$$

and we obtain $R_y^2 - \gamma^2 = \left(1-\gamma^2\right)/(1+8\sigma^2)$, from which it follows that

$$\sigma^2 = \frac{1 - R_y^2}{8\left(R_y^2 - \gamma^2\right)}.$$

For $m = 0$, $\sigma^2 = (1 - R_y^2)/5\left(R_y^2 - \gamma^2\right)$.

In the unit root case, using results in Section A.3 of the Appendix together with the above we have

$$\begin{aligned} R_{\Delta y}^2 &= \frac{\beta^2}{\beta^2 + \mu_\delta^2 + \sum_{\ell=1}^m \left(N^{-1}\sum_{i=1}^N c_{\ell i}^2\right) + \sigma^2} \\ &= \frac{1}{1 + \mu_\delta^2 + \sum_{\ell=1}^m \left(\sigma_{\vartheta\ell}^2 + \mu_{\vartheta\ell}^2\right) + \sum_{\ell=1}^m \sigma_{\eta\ell}^2 + \sigma^2}. \end{aligned}$$

As in the stationary case we control the value of $R_{\Delta y}^2$ to be the same for all values of m. Using (54) this leads to

$$\sigma^2 = \frac{1 - R_{\Delta y}^2}{8R_{\Delta y}^2} \text{ for } m \neq 0 \text{ and } \sigma^2 = \frac{1 - R_{\Delta y}^2}{5R_{\Delta y}^2} \text{ for } m = 0.$$

The parameter σ^2 is set such that $R_y^2 = 0.8$ and $R_{\Delta y}^2 = 0.4$, for all values of m. In line with the above derivations we set $\beta = 1$ and for γ we consider the values $\gamma = \{0.4, 0.8, 1.0\}$.

We consider the same combinations of T and N as in the AR(1) case, namely $T = \{5, 10\}$ and $N = \{100, 300, 500\}$ and report simulation results for the same set of statistics, for both γ and β, including size and power. Power is presented for $\gamma = \{0.38,\ 0.78\}$ and $\beta = 0.98$ for the null values of $\gamma = \{0.4, 0.8\}$ and $\beta = 1$, and for $\gamma = 0.98$ and $\beta = 0.95$ for the null values of $\gamma = 1$ and $\beta = 1$. As previously, all tests are carried out at the 5% significance level and all experiments are replicated 2,000 times, unless otherwise stated. The standard errors used for inference are based on the same formulas as those used in the AR(1) case with all derivatives computed numerically.

8.2 Monte Carlo results

We begin by reporting on the performance of the sequential multiple testing LR (MTLR) procedure for selecting the true number of factors. We consider the performance of the QML estimator when the number of factors is estimated using the MTLR procedure as well as when the number of factors is set to its true value, m_0. For this set of experiments the fixed effects are allowed to be correlated with the errors, and with the regressors in the panel ARX case. In the above Monte Carlo designs this corresponds to setting $b_1 = b_2 = 1$, with the additional b_0 parameter set to 1 for the ARX(1) model. We conclude this section by presenting results for the QML estimator together with the GMM quasi-difference (QD) and first-difference (FD) estimators of ALS, when the number of factors is assumed to be known. In this set of experiments the fixed effects are not correlated with the errors, as this would render the GMM estimators inconsistent. This corresponds to setting $b_1 = 0$ and $b_2 = 1$, with the additional b_0 parameter set to 1 for the ARX(1) model. However, fixed effects are allowed to be correlated with the regressors in the case of the ARX(1) design.

Results for the unit root case provided in Section S.5 of the online supplement in Tables S11-S19, show that the sequential MTLR procedure works very well even in the unit root case. The performance is very similar to the stationary case with $|\gamma| < 1$, and indeed, the probability of selecting the true number of factors exceeds 95% in most cases even under non-Gaussianity. Furthermore, for the AR(1) model both the bias and RMSE are sufficiently small and the empirical size is close to the nominal level regardless of whether the number of factors is estimated or not, and the error term is Gaussian or not. Similar results are found for the ARX(1) model for which the bias and RMSE are small and inference is accurate with reasonably high power regardless of whether the number of factors is estimated or not, and the error term is Gaussian or not.

8.2.1 Selection of the number of common factors

Tables 1 and 2 provide results on the performance of the sequential MTLR procedure for the AR(1) and the ARX(1) models, respectively. Specifically they report the number of times, in percent, that the estimated number of factors, $\hat{m}$, based on the sequential MTLR procedure outlined in Section 7.1 is equal to the true number of factors m_0. The sequential MTLR procedure is implemented using the $\mathcal{LR}_N\left(m_{\max}, m_0\right)$ statistic for testing $m = m_0 = \{0, 1, 2, .., T-3\}$ against $m = m_{\max} = T-2$, with significance level $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$, using the critical values of the chi-square distribution with degrees of freedom given by (44). Results are reported for the case of both Gaussian and non-Gaussian errors. The tables show that the estimator $\hat{m}$ performs very well. Even for the case where $N = 100$, the true number of factors is estimated quite precisely. We also find that as N gets larger, the probability of selecting the true number of factors approaches 100%, which supports the consistency of the proposed procedure.

Table 1: Empirical frequency of correctly selecting the true number of factors, m_0, using the sequential MTLR procedure in the case of the AR(1) model

N	$T=5$						$T=10$					
	$m_0=0$		$m_0=1$		$m_0=2$		$m_0=0$		$m_0=1$		$m_0=2$	
	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian
$\gamma = 0.4$												
100	99.2	95.8	98.8	95.7	98.3	97.6	99.3	95.8	99.0	95.1	99.5	96.9
300	99.8	97.9	99.7	98.4	100.0	99.1	99.9	97.6	99.8	98.5	100.0	99.5
500	99.9	97.7	100.0	98.6	99.9	99.4	99.9	98.9	99.9	99.0	100.0	99.3
1,000	99.9	98.9	100.0	99.4	100.0	99.5	100.0	99.3	100.0	99.6	99.9	99.8
$\gamma = 0.8$												
100	98.9	96.1	99.0	94.9	98.6	96.9	99.2	96.0	99.1	96.3	99.6	96.1
300	99.8	97.0	99.5	98.3	99.7	99.0	99.8	97.6	99.9	99.0	100.0	98.9
500	99.9	97.7	99.9	98.2	99.6	98.9	100.0	97.9	99.7	99.3	100.0	98.8
1,000	99.9	98.7	100.0	99.2	99.9	99.1	100.0	99.2	100.0	99.5	100.0	99.5

Note: y_{it} is generated as $y_{it} = \alpha_i + \mu_\delta \delta_t + \gamma y_{i,t-1} + \zeta_{it}$, $\zeta_{it} = \boldsymbol{\eta}_i' \mathbf{f}_t + u_{it}$, for $i = 1, 2, ..., N; t = -49, 48, ...0, 1, ..., T$, with $y_{i,-49} = \frac{\alpha_i}{1-\gamma} + \sum_{j=0}^{49} \gamma^j \zeta_{i,-j}$. The idiosyncratic errors are generated as $u_{it} \sim IID\mathcal{N}(0, \sigma^2)$ under Gaussianity and $u_{it} \sim IID\frac{\sigma}{\sqrt{12}}(\chi_6^2 - 6)$ under non-Gaussianity where χ_6^2 is a chi-square variate with 6 degrees of freedom and $\sigma^2 = 1$. The fixed effects are generated as $\alpha_i = \bar{u}_i + v_i$, where $\bar{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$ and $v_i \sim IID\mathcal{N}(0, 1)$. The remaining parameters are generated as described in Section 8.1.1. Each f_t is generated once and the same $f_t's$ are used throughout the replications. The first 49 observations are discarded. $\hat{m}$ is the estimated number of factors computed using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$. All experiments are based on 1,000 replications.

Table 2: Empirical frequency of correctly selecting the true number of factors, m_0, using the sequential MTLR procedure in the case of the ARX(1) model

N	$T=5$						$T=10$					
	$m_0=0$		$m_0=1$		$m_0=2$		$m_0=0$		$m_0=1$		$m_0=2$	
	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian
$\gamma = 0.4$												
100	98.8	93.7	98.6	95.9	97.5	96.9	99.1	94.4	99.3	95.2	99.0	94.9
300	99.5	97.3	99.4	98.2	99.6	99.1	99.7	97.7	99.7	98.6	99.6	98.4
500	99.9	99.0	99.7	99.0	99.8	99.1	100.0	98.3	100.0	98.7	100.0	99.4
1,000	99.7	99.0	100.0	99.1	99.9	99.5	100.0	99.6	99.8	99.8	100.0	99.2
$\gamma = 0.8$												
100	98.4	94.8	98.4	96.7	98.1	96.8	98.3	93.9	99.2	93.7	99.1	95.2
300	99.9	97.6	99.6	97.9	99.6	99.3	99.8	98.1	99.9	98.1	99.9	98.1
500	99.9	98.3	99.9	99.2	99.9	99.6	99.8	99.0	100.0	99.1	99.9	98.9
1,000	99.7	99.0	99.9	99.3	99.9	99.7	99.9	99.1	99.9	99.7	100.0	99.5

Note: y_{it} is generated as $y_{it} = \alpha_i + \mu_\delta \delta_t + \gamma y_{i,t-1} + \beta x_{it} + \zeta_{it}$, $\zeta_{it} = \boldsymbol{\eta}_i' \mathbf{f}_t + u_{it}$ for $i = 1, 2, ..., N; t = -49, 48, ...0, 1, ..., T$, with $y_{i,-49} = \frac{\alpha_i}{1-\gamma} + \beta \sum_{j=0}^{49} \gamma^j x_{i,-j} + \sum_{j=0}^{49} \gamma^j \zeta_{i,-j}$, and $\beta = 1$. The idiosyncratic errors are generated as $u_{it} \sim IIDN(0, \sigma^2)$ under Gaussianity and $u_{it} \sim IID\frac{\sigma}{\sqrt{12}}(\chi_6^2 - 6)$ under non-Gaussianity where χ_6^2 is a chi-square variate with 6 degrees of freedom and $\sigma^2 = (1 - R_y^2)/8 \left(R_y^2 - \gamma^2\right)$ with $R_y^2 = 0.8$. The fixed effects, α_i, are generated as $\alpha_i = \bar{x}_i + \bar{u}_i + v_i$, where $\bar{x}_i = T^{-1} \sum_{t=1}^{T} x_{it}$, $\bar{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$ and $v_i \sim IID\mathcal{N}(0, 1)$. The remaining parameters are generated as described in Section 8.1.2. When $m_0 = 0$, $\zeta_{it} = u_{it}$ and $\sigma^2 = (1 - R_y^2)/5 \left(R_y^2 - \gamma^2\right)$. See also the notes to Table 1.

8.2.2 Performance of the QML estimator

The previous Monte Carlo simulation results reveal that the sequential MTLR procedure performs very well in selecting the true number of unobserved factors. We next consider the performance of the proposed estimator when the number of factors is estimated based on this procedure. The results for the case where the number of factors is known are also included for comparison. Results are reported for the case of non-Gaussian errors. The corresponding results for Gaussian errors are provided in the online supplement.

AR(1) For this case the bias and RMSE, both multiplied by 100, are reported in Table 3, with Table 4 providing associated empirical size and power for the QML estimates of γ. The number of factors, when estimated, is computed based on the sequential MTLR procedure described in Section 7.1 with the significance level $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$. The results show that the effects of estimating the number of factors is negligible in all cases. The biases, RMSEs, sizes and powers with the true and estimated number of factors are very similar. The overall performance of the bias and RMSE is favourable except for the case where $T = 5$, $N = 100$ and $\gamma = 0.8$. In this case, the bias and RMSE are relatively large. However, the results improve with N, as predicted by the asymptotic theory. Similarly, we find that the test size and power are satisfactory expect when T is small and γ relatively large. For example, in the case of experiments with $T = 5$ and $\gamma = 0.8$ there is some evidence of size distortion when $N \leq 500$, although the size distortion reduces as N and T are increased. See Table A1 in the Appendix where we also provide results for $N = 1,000$ and $2,000$.

Table 3: Bias(×100) and RMSE(×100) of γ for the AR(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	$T=5$								$T=10$							
	Bias (×100)				RMSE (×100)				Bias (×100)				RMSE (×100)			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
$\gamma = 0.4$																
100	0.47	-0.14	-0.63	-1.55	9.26	9.36	13.19	14.64	-0.13	-0.12	-0.11	-0.29	4.30	4.39	4.31	4.54
300	0.18	0.00	-0.21	-0.50	4.89	5.00	7.02	7.93	-0.02	-0.05	-0.08	-0.13	2.37	2.35	2.56	2.54
500	0.13	-0.10	-0.15	-0.17	3.68	3.91	5.22	5.39	-0.05	-0.05	0.00	-0.03	1.83	1.83	1.90	1.91
$\gamma = 0.8$																
100	0.35	-0.68	-12.14	-13.51	12.42	14.98	30.09	31.68	0.54	0.44	0.47	0.08	6.19	6.67	6.85	7.11
300	1.30	1.21	-2.00	-2.67	9.47	10.27	16.04	16.85	0.17	0.03	0.10	0.01	3.34	3.55	3.75	3.72
500	1.45	1.22	-0.35	-0.39	8.12	8.40	12.13	12.47	0.03	0.05	0.06	0.00	2.35	2.53	2.60	2.67

Note: $\hat{m}$ is estimated using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$; γ is the coefficient of the lagged dependent variable given in (1) in the absence of the $\mathbf{x}_{it}$ regressors. All experiments are based on 2,000 replications. See also the notes to Table 1.

Table 4: Size(%) and power(%) of γ for the AR(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	$T=5$								$T=10$							
	Size				Power				Size				Power			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
	H_0: $\gamma = 0.4$				H_1: $\gamma = 0.3$				H_0: $\gamma = 0.4$				H_1: $\gamma = 0.3$			
100	6.1	6.2	5.6	5.7	27.3	26.9	15.2	17.3	6.7	6.2	4.4	5.0	68.7	65.7	61.0	60.7
300	5.5	4.6	4.7	5.7	55.9	54.9	32.7	32.6	5.1	5.2	5.0	4.6	98.6	98.2	97.8	97.4
500	4.9	4.8	4.2	4.9	74.5	76.1	48.2	46.8	5.0	4.5	4.4	4.7	99.9	99.9	99.9	100.0
	H_0: $\gamma = 0.8$				H_1: $\gamma = 0.7$				H_0: $\gamma = 0.8$				H_1: $\gamma = 0.7$			
100	23.3	21.8	26.0	28.5	24.2	25.1	28.2	30.8	11.6	11.3	9.3	9.4	54.5	53.1	42.1	44.5
300	19.0	19.3	15.8	14.7	32.4	30.2	21.6	20.2	5.8	5.7	5.2	5.1	85.0	85.3	78.7	77.1
500	16.9	18.4	11.7	12.4	36.3	39.6	21.3	22.9	5.0	5.2	4.0	4.5	96.1	95.6	93.0	92.6

See the notes to Table 3.

ARX(1) Simulation results for the ARX(1) model are provided in Tables 5 and 6. Similar results as in the AR(1) model are found for the ARX(1). Comparing the bias and RMSE values of the γ and β estimators for the case of the true and estimated number of factors, these appear to be very similar and are also very small. With regard to size and power, unlike the AR(1) model, the empirical sizes are close to the nominal level in all cases and power is reasonably high even when the number of factors is estimated.

Table 5: Bias(×100) and RMSE(×100) of γ and β for the ARX(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	$T=5$								$T=10$							
	Bias (×100)				RMSE (×100)				Bias (×100)				RMSE (×100)			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
γ, $\gamma=0.4$																
100	0.03	-0.04	-0.04	0.00	1.44	1.47	1.54	1.58	-0.03	-0.01	-0.04	-0.01	0.82	0.80	0.88	0.85
300	-0.03	-0.01	0.04	-0.02	0.84	0.83	0.90	0.91	-0.01	0.00	0.01	-0.01	0.46	0.45	0.49	0.49
500	0.01	-0.01	-0.02	0.01	0.65	0.64	0.66	0.69	0.00	0.00	0.00	-0.01	0.36	0.36	0.38	0.38
β																
100	-0.08	0.04	0.06	-0.06	1.91	1.90	2.03	2.09	0.05	-0.01	0.02	0.02	1.12	1.15	1.18	1.20
300	-0.02	0.02	-0.05	0.07	1.05	1.07	1.17	1.15	0.02	0.01	0.00	0.00	0.65	0.65	0.69	0.69
500	0.00	0.00	0.02	-0.01	0.83	0.84	0.88	0.90	0.01	0.00	0.00	0.02	0.51	0.50	0.54	0.54
γ, $\gamma=0.8$																
100	0.07	-0.04	0.02	0.02	1.88	1.99	1.89	1.94	-0.04	-0.05	-0.03	-0.01	0.83	0.84	0.85	0.86
300	-0.05	-0.02	0.03	-0.01	1.09	1.10	1.08	1.08	-0.01	0.01	0.02	-0.02	0.47	0.47	0.48	0.48
500	0.01	-0.02	-0.03	0.01	0.84	0.84	0.81	0.84	-0.01	-0.01	-0.01	0.00	0.36	0.36	0.36	0.37
β																
100	-0.13	-0.03	0.07	-0.06	3.47	3.47	3.58	3.64	0.07	-0.05	0.01	0.12	1.98	2.00	2.00	2.14
300	-0.05	-0.03	-0.05	0.12	1.93	1.98	2.05	2.00	0.02	0.01	0.00	0.02	1.14	1.17	1.19	1.16
500	0.01	0.00	0.02	-0.01	1.52	1.49	1.57	1.57	0.01	0.00	0.00	0.01	0.90	0.88	0.92	0.92

Note: $\hat{m}$ is estimated using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$; γ and β are the coefficients of the lagged dependent variable and the $\mathbf{x}_{it}$ regressor given in (1). All experiments are based on 2,000 replications. See also the notes to Table 2.

Table 6: Size(%) and power(%) of γ and β for the ARX(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	$T=5$								$T=10$							
	Size				Power				Size				Power			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
γ	H_0: $\gamma=0.40$				H_1: $\gamma=0.38$				H_0: $\gamma=0.40$				H_1: $\gamma=0.38$			
100	5.3	6.4	5.6	7.4	29.4	30.7	27.9	27.0	6.3	5.7	6.2	6.1	73.3	71.1	67.1	67.2
300	5.6	5.4	6.1	6.0	68.5	67.4	60.7	62.8	5.5	4.3	5.5	5.4	99.3	99.3	98.3	98.2
500	5.9	4.5	5.0	5.7	86.6	87.8	85.1	82.4	5.4	5.3	4.9	5.0	100.0	100.0	100.0	100.0
β	H_0: $\beta=1$				H_1: $\beta=0.98$				H_0: $\beta=1$				H_1: $\beta=0.98$			
100	6.1	6.5	6.6	6.5	21.2	20.5	18.0	19.5	5.0	5.7	5.5	5.6	42.0	43.7	39.6	39.1
300	4.6	4.7	5.1	5.5	45.9	46.2	43.8	39.8	5.1	4.9	5.2	5.6	86.0	86.0	82.5	83.8
500	5.2	4.6	5.2	4.9	66.5	67.3	60.4	62.2	5.3	4.6	5.9	6.7	97.5	98.3	96.1	95.7
γ	H_0: $\gamma=0.80$				H_1: $\gamma=0.78$				H_0: $\gamma=0.80$				H_1: $\gamma=0.38$			
100	5.5	7.0	5.7	6.9	20.0	22.7	20.2	21.4	6.5	6.4	7.2	6.0	69.0	69.9	69.2	67.1
300	5.8	5.2	5.6	5.2	48.8	46.2	47.1	48.9	5.7	5.1	5.7	5.3	99.3	98.6	98.4	98.6
500	4.8	5.0	5.2	5.6	66.1	67.9	70.4	66.7	5.0	4.9	5.0	5.1	100.0	100.0	99.9	99.9
β	H_0: $\beta=1$				H_1: $\beta=0.98$				H_0: $\beta=1$				H_1: $\beta=0.98$			
100	6.1	6.2	6.9	6.1	10.9	10.7	10.0	11.2	4.9	5.5	5.5	6.6	15.9	18.7	16.3	16.1
300	4.7	4.7	5.7	5.2	18.0	17.5	19.2	14.9	4.7	5.5	5.4	4.7	41.6	40.7	40.4	38.5
500	5.3	4.1	5.0	6.0	26.2	25.6	24.4	25.3	5.9	5.1	5.5	5.6	61.8	60.5	59.8	58.8

See the notes to Table 5.

8.2.3 QML and GMM results

Next we present simulation results comparing our QML estimator with the GMM estimator of ALS in the case of non-Gaussian errors. Corresponding results for the case of Gaussian errors are available in the online supplement. For this set of experiments the number of factors during estimation is set to the true number of factors. The GMM estimators include the quasi-difference and first- difference ALS one step and two step estimators denoted by QD1, QD2, FD1 and FD2, respectively, computed as detailed in Section S.3 of the online supplement. Recall that for these results the individual fixed effects are not correlated with the errors, only with the regressor in the case of the ARX(1) model. The results for the AR(1) model are summarised in Tables 7 and 8, and for the ARX(1) model are summarised in Tables 9 and 10.

AR(1) From Table 7, we find that the QML estimator performs (sometimes substantially) better than the GMM estimators in terms of bias and RMSE.[4] This is particularly evident for $\gamma = 0.8$. When $\gamma = 0.8$, the GMM estimators, especially FD1 and FD2, perform very poorly due to weak instruments whereas the QML estimator has small bias and RMSE. With regard to size and power shown in Table 8, the GMM estimators have substantial size distortions while the QML estimator has empirical sizes close to the nominal value except for the case with $\gamma = 0.8$ and $N = 100$, as in Table 3.

Table 7: Bias($\times$100) and RMSE($\times$100) of γ for the QML and GMM estimators in the case of the AR(1) model, using the true number of factors, m_0

N	$T = 10$																			
	Bias ($\times$100)					RMSE ($\times$100)					Bias ($\times$100)					RMSE ($\times$100)				
	QML	GMM				QML	GMM				QML	GMM				QML	GMM			
		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2
m_0	1										2									
$\gamma = 0.4$																				
100	-0.25	11.28	14.82	-34.26	-23.66	4.03	36.77	36.08	35.68	25.81	-0.08	49.13	49.08	-24.60	-23.11	4.43	52.52	53.20	56.20	49.08
300	0.05	-0.58	4.86	-18.35	-7.62	2.43	24.26	22.88	19.52	9.40	-0.01	48.90	48.03	-13.13	-12.50	2.52	52.57	52.60	43.70	35.49
500	-0.03	-1.73	3.18	-12.92	-4.15	1.85	19.57	18.62	14.00	5.74	-0.02	47.35	46.34	-7.83	-8.47	1.96	51.75	51.70	37.19	29.68
$\gamma = 0.8$																				
100	0.33	12.15	12.09	-70.73	-62.84	6.10	21.40	22.05	72.22	65.63	0.56	17.61	17.60	-70.87	-67.90	7.01	18.87	18.84	90.22	87.55
300	0.32	10.06	10.85	-51.42	-35.05	3.55	19.88	19.35	52.57	37.34	0.14	17.36	17.22	-50.84	-43.89	3.79	18.53	18.46	71.89	64.92
500	0.00	8.40	9.54	-42.67	-24.92	2.41	19.00	18.01	43.70	26.93	0.10	16.83	16.91	-43.55	-35.47	2.83	18.41	18.23	64.29	56.47

Note: GMM QD1, QD2, FD1 and FD2 are the quasi-difference and first- difference ALS one step and two step estimators respectively computed as described in Section S.3 of the supplementary material. All experiments are based on 2,000 replications. See also the notes to Tables 1 and 3.

Table 8: Size(%) and power(%) of γ for the QML and GMM estimators in the case of the AR(1) model, using the true number of factors, m_0

N	$T = 10$																			
	Size					Power					Size					Power				
	QML	GMM				QML	GMM				QML	GMM				QML	GMM			
		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2
m_0	1										2									
γ	H_0: $\gamma = 0.40$					H_1: $\gamma = 0.34$					H_0: $\gamma = 0.40$					H_1: $\gamma = 0.34$				
100	5.9	64.7	73.1	89.2	89.9	34.1	78.0	85.9	96.6	97.7	5.5	86.4	87.7	29.0	41.7	29.3	85.6	87.4	29.9	43.3
300	5.7	46.5	38.9	72.6	48.3	70.3	71.7	70.1	92.4	86.7	4.5	83.8	84.9	20.1	32.5	65.6	83.6	85.1	19.6	32.4
500	6.0	38.2	26.6	59.5	29.8	90.8	72.3	72.1	90.1	83.4	5.3	80.4	81.5	15.1	27.9	86.9	80.1	82.0	14.9	28.4
γ	H_0: $\gamma = 0.80$					H_1: $\gamma = 0.74$					H_0: $\gamma = 0.80$					H_1: $\gamma = 0.74$				
100	10.9	93.6	96.7	99.6	99.8	32.0	95.1	97.6	99.8	100.0	10.3	96.5	97.2	51.1	64.6	26.2	94.6	96.6	53.3	73.7
300	6.5	88.0	89.9	98.4	96.2	51.6	92.1	92.9	99.2	99.1	4.8	95.0	95.7	36.4	47.7	44.6	93.9	95.3	36.9	62.2
500	4.6	83.2	82.8	97.3	91.6	71.8	88.8	88.6	99.3	98.2	5.1	93.0	94.0	31.0	43.2	59.8	91.5	93.7	31.8	60.6

See the notes to Table 7.

[4] The case of $T = 5$ is not reported for the AR(1) model because the number of unknown parameters exceeds that of the moment conditions.

ARX(1) Compared to the pure AR(1) case, the performance of all the estimators is improved for the ARX(1) model as shown in Table 9. This is especially so for FD1 and FD2. However, in terms of relative performance, the QML estimator still outperforms the GMM estimators with regard to bias and RMSE. With regard to size and power, reported in Table 10, the QML estimator has empirical sizes close to the nominal level for all combinations including $N = 100$ and $\gamma = 0.8$ for which some size distortions were reported for the pure AR(1) case. For the GMM estimators, the performance crucially depends on γ, m_0, N and T, and there is no GMM estimator that performs well for all combinations, which is in contrast to the QML estimator that performs well for all cases considered. For instance, when $T = 5$, QD1 and FD1 tend to have correct empirical sizes except for $\gamma = 0.8$. However, they tend to have large size distortions when T is increased to $T = 10$ for $m_0 = 1$. QD2 and FD2 tend to have larger size distortions than QD1 and FD1.[5]

[5] Since both QD2 and FD2 are nonlinear GMM estimators, it is not straightforward to apply the Windmeijer (2005) correction.

Table 9: Bias(×100) and RMSE(×100) of γ and β for the QML and GMM estimators in the case of the ARX(1) model, using the true number of factors, m_0

N	$T=5$										$T=10$									
	Bias (×100)					RMSE (×100)					Bias (×100)					RMSE (×100)				
	QML	GMM				QML	GMM				QML	GMM				QML	GMM			
		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2
m_0 1																				
γ, $\gamma=0.4$																				
100	0.03	0.38	0.08	0.08	0.14	1.44	3.20	2.72	4.04	2.89	-0.03	10.54	10.46	-2.46	-2.12	0.82	21.05	20.96	3.33	2.90
300	-0.03	0.43	0.12	0.03	0.06	0.84	1.41	1.34	2.34	1.61	-0.01	1.92	1.67	-0.87	-0.37	0.46	8.00	7.66	1.83	0.82
500	0.01	0.50	0.17	0.04	0.06	0.65	1.13	1.03	1.96	1.28	0.00	0.81	0.57	-0.56	-0.17	0.36	3.51	3.20	1.26	0.54
β																				
100	-0.08	-0.19	-0.04	-0.06	-0.08	1.91	2.70	2.32	4.54	3.28	0.05	-9.23	-9.13	-0.68	-0.53	1.12	20.30	20.11	3.91	3.38
300	-0.02	-0.11	-0.04	0.13	0.06	1.05	1.31	1.32	2.46	1.79	0.02	-1.40	-1.29	-0.10	0.04	0.65	7.90	7.29	2.71	1.13
500	0.00	-0.12	-0.03	0.16	0.06	0.83	1.03	1.02	2.12	1.46	0.01	-0.39	-0.30	0.03	0.05	0.51	3.46	3.04	2.14	0.74
γ, $\gamma=0.8$																				
100	0.07	8.56	7.19	-0.58	-0.43	1.88	10.14	9.32	5.49	4.58	-0.04	12.52	12.45	-5.83	-5.05	0.83	12.55	12.48	6.55	5.73
300	-0.05	8.30	5.24	-0.04	-0.07	1.09	9.62	7.61	3.02	2.42	-0.01	12.28	11.79	-2.04	-0.81	0.47	12.29	11.80	3.11	1.26
500	0.01	8.63	4.87	0.07	0.03	0.84	9.72	7.18	2.39	1.88	-0.01	12.23	11.65	-1.20	-0.34	0.36	12.24	11.66	2.02	0.73
β																				
100	-0.13	-1.87	-0.41	-0.52	-0.35	3.47	6.97	5.31	7.83	6.36	0.07	-14.40	-13.87	-4.45	-3.82	1.98	15.70	15.20	9.78	8.50
300	-0.05	-0.26	-0.05	-0.13	-0.04	1.93	3.16	2.81	4.25	3.37	0.02	-12.45	-8.73	-1.22	-0.39	1.14	12.99	9.37	5.72	2.39
500	0.01	-0.03	0.02	-0.14	-0.03	1.52	2.24	2.12	3.32	2.59	0.01	-12.00	-7.62	-0.46	-0.07	0.90	12.33	8.05	4.50	1.48
m_0 2																				
γ, $\gamma=0.4$																				
100	-0.04	0.19	0.12	-2.08	-1.69	1.54	4.17	3.67	7.19	6.71	-0.04	-0.30	-0.22	-1.21	-0.97	0.88	5.20	5.06	4.58	3.43
300	0.04	0.03	-0.04	-0.56	-0.41	0.90	1.67	1.65	3.73	3.33	0.01	-0.33	-0.17	-0.28	-0.10	0.49	1.64	1.16	4.90	1.41
500	-0.02	-0.11	-0.16	-0.21	-0.16	0.66	1.27	1.27	2.60	2.25	0.00	-0.25	-0.11	0.09	0.00	0.38	1.21	0.64	3.60	0.91
β																				
100	0.06	-0.06	0.16	-2.21	-1.95	2.03	3.27	3.23	12.44	12.42	0.02	0.23	0.19	-0.04	-0.01	1.18	4.65	4.43	5.21	4.01
300	-0.05	-0.07	0.01	-0.38	-0.15	1.17	1.80	1.80	6.30	5.94	0.00	0.26	0.14	0.21	0.03	0.69	1.31	1.05	4.43	1.47
500	0.02	0.04	0.11	-0.24	-0.04	0.88	1.37	1.33	4.57	4.21	0.00	0.15	0.05	-0.04	0.00	0.54	1.05	0.79	3.49	1.07
γ, $\gamma=0.8$																				
100	0.02	7.75	5.63	-18.67	-18.03	1.89	9.59	8.30	32.53	33.10	-0.03	-0.67	-0.57	-3.62	-2.92	0.85	3.76	3.64	6.94	5.38
300	0.03	4.54	2.68	-7.99	-7.75	1.08	7.19	5.66	18.74	19.62	0.02	-1.10	-0.74	-1.42	-0.70	0.48	1.47	1.15	4.44	1.83
500	-0.03	2.67	1.55	-4.05	-3.90	0.81	5.67	4.31	13.15	13.79	-0.01	-0.93	-0.57	-0.77	-0.34	0.36	1.24	0.88	3.72	1.05
β																				
100	0.07	0.04	0.50	-21.24	-20.71	3.58	7.26	6.89	38.94	39.69	0.01	0.63	0.60	-0.94	-0.62	2.00	3.79	3.59	9.92	7.64
300	-0.05	0.50	0.12	-10.05	-9.48	2.05	3.98	3.48	23.11	23.68	0.00	0.57	0.34	-0.05	-0.03	1.19	2.00	1.73	6.11	2.69
500	0.02	0.50	0.16	-5.35	-4.82	1.57	2.99	2.57	16.65	16.97	0.00	0.42	0.22	-0.29	0.01	0.92	1.64	1.35	5.21	1.81

Note: GMM QD1, QD2, FD1 and FD2 are the quasi-difference and first- difference ALS one step and two step estimators respectively computed as described in Section S.3 of the supplementary material. All experiments are based on 2,000 replications. See also the notes to Tables 2 and 5.

Table 10: Size(%) and power(%) of γ and β for the QML and GMM estimators in the case of the ARX(1) model, using the true number of factors, m_0

N	$T=5$										$T=10$									
	Size					Power					Size					Power				
	QML	GMM				QML	GMM				QML	GMM				QML	GMM			
		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2
m_0 1																				
γ	H_0: $\gamma=0.40$					H_1: $\gamma=0.38$					H_0: $\gamma=0.40$					H_1: $\gamma=0.38$				
100	5.3	6.8	15.4	4.7	13.8	29.5	14.3	30.0	9.2	22.6	6.3	31.0	88.0	28.1	83.2	73.1	33.2	88.1	68.3	98.2
300	5.6	7.0	11.1	4.5	8.4	68.5	24.8	42.6	16.8	30.2	5.5	17.2	47.3	13.1	25.5	99.3	28.0	78.1	74.4	98.2
500	5.9	7.8	9.5	4.3	7.2	86.6	32.0	55.6	24.3	43.1	5.4	16.5	38.3	9.9	15.6	100.0	35.8	88.5	85.0	99.9
β	H_0: $\beta=1$					H_1: $\beta=0.98$					H_0: $\beta=1$					H_1: $\beta=0.98$				
100	6.1	5.9	15.0	3.8	12.9	21.2	17.1	29.7	8.3	21.4	5.0	28.5	86.0	6.6	73.7	42.0	33.2	88.5	14.4	80.6
300	4.6	4.6	9.8	4.6	6.8	46.0	34.6	45.4	12.5	25.3	5.1	10.9	36.4	5.2	27.0	86.0	30.1	80.7	14.9	74.7
500	5.2	4.7	8.9	4.2	7.5	66.5	53.9	62.1	18.1	34.9	5.2	9.1	26.5	4.5	15.8	97.5	43.6	91.7	18.0	89.0
γ	H_0: $\gamma=0.80$					H_1: $\gamma=0.78$					H_0: $\gamma=0.80$					H_1: $\gamma=0.78$				
100	5.5	77.3	71.6	5.8	14.0	20.0	78.8	73.4	9.2	20.6	6.6	100.0	100.0	54.9	95.4	68.8	100.0	100.0	78.0	99.2
300	5.8	76.1	56.3	5.0	8.5	48.8	79.0	56.6	10.8	20.9	5.7	100.0	100.0	21.0	41.0	99.3	100.0	100.0	59.7	96.9
500	4.8	78.6	54.1	4.0	7.6	66.1	80.9	55.4	13.7	25.6	5.0	100.0	100.0	13.3	21.3	100.0	100.0	100.0	59.1	98.8
β	H_0: $\beta=1$					H_1: $\beta=0.98$					H_0: $\beta=1$					H_1: $\beta=0.98$				
100	6.1	11.2	17.2	4.9	13.4	10.9	14.9	20.4	6.3	16.9	5.0	85.1	98.3	15.5	80.2	15.8	90.5	99.4	20.8	84.0
300	4.7	6.6	12.3	5.3	8.9	18.1	12.6	20.9	7.7	15.0	4.6	98.5	97.4	7.7	32.3	41.5	99.9	99.3	13.0	54.1
500	5.3	4.7	10.1	3.8	7.5	26.2	14.7	24.6	9.4	16.5	5.9	100.0	99.0	5.9	18.7	61.9	100.0	99.9	12.0	54.5
m_0 2																				
γ	H_0: $\gamma=0.40$					H_1: $\gamma=0.38$					H_0: $\gamma=0.40$					H_1: $\gamma=0.38$				
100	5.6	5.7	11.6	7.8	13.1	27.9	11.5	21.0	13.2	19.1	6.2	11.9	58.2	8.8	60.1	67.0	51.1	90.8	28.3	80.9
300	6.1	5.1	8.8	5.3	7.8	60.7	21.5	31.9	14.6	26.0	5.5	6.9	19.4	5.5	25.9	98.3	76.1	95.2	23.6	83.9
500	5.0	4.3	7.8	4.8	6.5	85.1	38.2	49.5	20.2	35.6	4.9	7.5	16.5	5.3	20.4	100.0	87.1	98.7	22.8	92.3
β	H_0: $\beta=1$					H_1: $\beta=0.98$					H_0: $\beta=1$					H_1: $\beta=0.98$				
100	6.5	5.2	11.6	6.1	10.6	18.0	10.6	19.4	8.3	15.7	5.5	6.9	52.2	8.4	61.1	39.5	12.2	58.9	11.0	66.5
300	5.1	5.0	8.5	4.5	7.1	43.9	21.3	29.2	9.6	18.0	5.2	4.9	20.3	4.9	24.8	82.6	28.0	68.4	11.0	62.6
500	5.2	4.7	7.9	4.5	7.4	60.5	30.9	38.0	14.3	25.2	5.9	5.0	16.4	4.3	18.7	96.0	45.5	85.0	11.5	77.7
γ	H_0: $\gamma=0.80$					H_1: $\gamma=0.78$					H_0: $\gamma=0.80$					H_1: $\gamma=0.78$				
100	5.7	73.4	55.8	31.8	36.5	20.2	73.9	57.1	34.9	38.7	7.1	24.6	70.5	21.2	75.5	69.3	68.2	94.7	37.6	89.3
300	5.6	45.4	30.8	20.5	22.7	47.1	49.0	40.3	24.4	32.2	5.7	20.9	35.2	8.9	36.2	98.4	89.6	97.1	25.0	87.5
500	5.2	30.5	22.7	14.5	14.6	70.4	43.3	40.4	20.4	31.7	5.0	25.2	30.8	6.7	27.8	99.9	95.9	99.4	20.2	94.1
β	H_0: $\beta=1$					H_1: $\beta=0.98$					H_0: $\beta=1$					H_1: $\beta=0.98$				
100	7.0	8.1	15.9	27.9	32.4	10.0	9.8	17.6	29.7	34.2	5.5	6.5	48.7	9.3	65.6	16.2	7.1	50.5	10.6	68.1
300	5.6	8.9	11.3	20.7	22.5	19.3	10.9	17.7	22.7	27.1	5.4	5.2	18.3	6.5	28.7	40.4	9.4	36.0	8.7	43.8
500	5.0	8.2	10.2	14.6	15.4	24.4	13.9	20.9	17.0	21.5	5.5	6.2	16.9	5.3	21.6	59.7	15.3	47.2	8.0	46.9

See the notes to Table 9.

9 An empirical application to growth convergence

In what follows, we apply the proposed QML approach to estimate panel growth regressions and evaluate unconditional convergence in economic growth across countries in the global economy. A number of studies have used basic cross section growth regressions for this purpose such as, for example, Barro (1991) and Mankiw et al. (1992), who examine a sample of 98 countries over the period 1960-1985, and Sala-i-Martin (1996) who considers 110 countries over the period 1960-1990, among others. The use of the basic cross section growth regression has received important criticisms by Islam (1995), Caselli et al. (1996) and by Lee et al. (1997,1998). Islam (1995) and Caselli et al. (1996) advocate and implement dynamic panel regressions including individual and time effects for studying growth convergence using five-yearly averages of growth rates as a way of abstracting from business cycle effects. In particular, Caselli et al. (1996) use first-differenced GMM estimators to deal with the fixed effects, but do not allow for interactive effects and implicitly assume that errors across countries are independent, which is unlikely to hold particularly considering the rapid increase in world trade and international financial linkages.

We estimate panel growth regressions using the log-level of real output at five year intervals, and the corresponding growth rate averaged over the same intervals, extending from 1960 to 2014. The last observation for the five year intervals is based on four years. The data is compiled from the latest version of the Penn World Tables by Heston, Summers and Aten (2012) and Feenstra et al. (2015). Output is measured by real GDP per capita constructed as the ratio of output-side real GDP at chained PPPs (in mil.2011US$) and population (in millions). We first estimate the number of factors using the proposed sequential MTLR procedure. Having selected the number of factors we then estimate two sets of panel regressions, one in the "level" of output per capita, and another in the "growth" rates, namely

$$\text{Levels:} \qquad y_{it} = \gamma y_{i,t-1} + \alpha_i + \delta_t + \boldsymbol{\eta}_i' \mathbf{f}_t + u_{it}, \text{ for } t = 1, 2, ..., T, \tag{55}$$

$$\text{Growth Rates:} \qquad \Delta y_{it} = \gamma_\Delta \Delta y_{i,t-1} + \alpha_{\Delta i} + \boldsymbol{\eta}_{\Delta i}' \mathbf{f}_{\Delta t} + \Delta u_{it}, \text{ for } t = 2, ..., T, \tag{56}$$

for $i = 1, 2, ..., N$, where y_{it} is the natural logarithm of real GDP per capita. Note the γ and γ_Δ are related but are not the same. For the levels regression y_{it} is measured at five year intervals, while for the growth rates regression y_{it} is measured as averages over five yearly intervals. We also report regression results for the case of no factors, for comparison. Note that the growth regression is not a first-differenced version of the level regression and has its own fixed and interactive time effects.

The QML estimates for the panel growth regressions, together with standard errors in parentheses, are presented in Table 18. The top panel of this table reports results for the level of the series, equation (55), and the bottom panel gives the results for the growth rates, equation (56). Starting with the top panel, the results show that the coefficients of the lagged dependent variable for the five year intervals have the correct signs. The estimated coefficient of the lagged dependent variable, $\hat{\gamma}$, for the level series without interactive effects, that is for $m = 0$, is equal to 0.967. Interestingly, this value is very close to that found by Lee et al. (1998) for their dynamic panel growth regressions including individual and time effects using the Summers-Heston data set over the period 1965-1989 with $N = 102$. This implies a speed of convergence of 0.007 based on the deterministic version of the Solow growth model where $\gamma = \exp(-\rho\tau)$ with ρ the speed of convergence and τ the time interval. Using the sequential MTLR procedure to select the number of factors yields $\hat{m} = 4$. The corresponding estimated value of $\hat{\gamma}$ now equals 0.918 with an implied speed of convergence of 0.017, which is much more plausible. These results show that inclusion of the unobserved factors in the level regression, (55), leads to a decrease in the persistence of $\hat{\gamma}$ and therefore an increase in the speed of convergence. Similar results are obtained for the growth rates regression summarised at the bottom panel of Table 18, where $\hat{\gamma}_\Delta$ drops from 0.288 for $m = 0$ to 0.150 for $\hat{m} = 3$, selected by the sequential MTLR procedure. These estimation results also confirm that persistence in the growth rates is rather small, irrespective of whether unobserved factors are allowed in the analysis. It is also of interest that the estimates of γ_Δ obtained with $m = 0$ are closer to the time series estimates obtained for individual countries that do not allow for possible common global effects.

Table 11: QML estimates of the panel growth AR(1) regressions over the period 1960-2014

	$\hat{\gamma}$	$\hat{\omega}$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_4$	$\hat{d}_5$	$\hat{d}_6$	$\hat{d}_7$	$\hat{d}_8$	$\hat{d}_9$	$\hat{d}_{10}$	$\hat{d}_{11}$	$\hat{d}_{12}$	$\hat{\sigma}^2$
Levels: y_{it} $(T = 11, N = 111)$														
$m = 0$	0.967	1.140	0.142	0.027	-0.018	-0.049	-0.069	0.046	0.016	0.060	0.039	0.002	-0.036	0.040
	(0.039)	(0.109)	(0.016)	(0.019)	(0.019)	(0.028)	(0.026)	(0.029)	(0.020)	(0.031)	(0.021)	(0.030)	(0.023)	
$\hat{m} = 4$	0.918	1.310	0.142	0.034	-0.011	-0.042	-0.065	0.046	0.019	0.063	0.045	0.010	-0.028	0.017
	(0.124)	(0.352)	(0.017)	(0.023)	(0.026)	(0.033)	(0.024)	(0.027)	(0.022)	(0.032)	(0.026)	(0.035)	(0.029)	
Growth Rates: Δy_{it} $(T = 10, N = 111)$														
$m = 0$	0.288	1.259	0.004	-0.006	-0.009	-0.011	0.013	0.000	0.011	0.004	-0.003	-0.002	-	0.002
	(0.064)	(0.128)	(0.004)	(0.004)	(0.006)	(0.006)	(0.007)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)		
$\hat{m} = 3$	0.150	1.706	0.004	-0.005	-0.010	-0.013	0.011	0.002	0.011	0.005	-0.002	-0.002	-	0.001
	(0.118)	(0.259)	(0.004)	(0.004)	(0.006)	(0.005)	(0.007)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)		

Note: T is the effective number of observations used in estimation and the figures in parentheses are standard errors. y_{it} and Δy_{it} are the natural logarithm of per capita GDP at five year intervals and the growth rate averaged over five year intervals, respectively. $\hat{m}$ is the estimated number of factors using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$.

10 Conclusion

This paper proposes a quasi maximum likelihood estimator for short dynamic panel data models with unobserved multiple common factors, where individual and time fixed effects are also explicitly included. This provides a natural extension of Hsiao, Pesaran, and Tahmiscioglu (2002) to panel data models with a multi-factor error structure. Our contribution can also be viewed as extending the standard dynamic panel data models with fixed and time effects, routinely used in the empirical literature, to allow for error cross sectional dependence through interactive time effects. We have also contributed to the literature on short T factor models where identification and estimation of the number of unobserved factors has proved to be challenging. In this regard, our proposed sequential multiple testing likelihood ratio (MTLR) procedure can be particularly relevant to the analysis of short T factor models. Monte Carlo results provide small sample evidence in support of the proposed QML estimator and show that the sequential MTLR procedure performs very well in selecting the number of unobserved factors. The same is also true for the performance of the QML estimators in terms of bias, RMSE and empirical size, and power. An empirical application to growth convergence using the most recent Penn World Tables suggests that allowing for interactive effects leads to estimates with a higher speed of convergence than previously indicated in the literature.

Although we allow the error variances to vary across units through the differences in factor loadings, it is assumed that the unit specific errors are cross sectionally homoskedastic, which is rather restrictive. However, our theoretical derivations can be readily adapted to cover the heteroskedastic error case, as was done in the recent paper by Hayakawa and Pesaran (2015) for models without unobserved common factors. It would also be interesting to extend the analysis to panel VAR models with interactive effects.

Appendix

A.1 Lemmas and their proofs

Lemma 1 *Consider the $T \times 1$ vector of composite errors $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, ...\xi_{iT})'$, where ξ_{i1} is defined by (10) and ξ_{it}, for $t = 2, 3, ..., T$ are defined by (6). Suppose that the conditions of Proposition 2 hold and T is fixed. Then*

$$\sup_i E \left\| \boldsymbol{\xi}_i \right\|^4 < K < \infty, \tag{A.1}$$

$$\sup_i E \left\| \mathbf{Z}_i \right\|^4 < K, \ \sup_i E \left\| \Delta \mathbf{y}_i \right\|^4 < K, \ \textit{and} \ \sup_i E \left\| \Delta \mathbf{W}_i \right\|^4 < K < \infty. \tag{A.2}$$

Proof. To obtain (A.1) note that

$$\left\| \boldsymbol{\xi}_i \right\|^4 = \left\| \boldsymbol{\xi}_i \boldsymbol{\xi}_i' \right\|^2 = Tr\left(\boldsymbol{\xi}_i \boldsymbol{\xi}_i' \boldsymbol{\xi}_i \boldsymbol{\xi}_i' \right) = \left(\boldsymbol{\xi}_i' \boldsymbol{\xi}_i \right)^2 = \left(\sum_{t=1}^{T} \xi_{it}^2 \right)^2 .$$

Then by Minkowski's inequality we have

$$E \left\| \boldsymbol{\xi}_i \right\|^4 = E \left(\sum_{t=1}^{T} \xi_{it}^2 \right)^2 \leq \left(\sum_{t=1}^{T} \left[E\left(\xi_{it}^4 \right) \right]^{1/2} \right)^2 ,$$

and since $\sup_i E(|\xi_{it}|^{4+\epsilon}) < K$ for $t = 1, 2, ..., T$ from result (15) of Proposition 2, result (A.1) follows noting that T is fixed. To establish (A.2), note that $\Delta \mathbf{W}_i = (\mathbf{I}_T, \mathbf{Z}_i, \Delta \widetilde{\mathbf{y}}_{i,-1}) = (\mathbf{I}_T, \mathbf{Z}_i, \mathbf{L}\Delta \mathbf{y}_i)$, where $\Delta \widetilde{\mathbf{y}}_{i,-1} = (0, \Delta y_{i1}, ..., \Delta y_{i,T-1})$, $\mathbf{Z}_i$ and $\Delta \mathbf{y}_i$ are given by (40) and (39), and

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & \cdots & \cdots & 0 \\ \vdots & 1 & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \tag{A.3}$$

with $\left\| \mathbf{L} \right\| < 1$. It is now easily seen that $\left\| \Delta \mathbf{W}_i \right\|^2 \leq T + \left\| \mathbf{Z}_i \right\|^2 + \left\| \Delta \mathbf{y}_i \right\|^2$, and by Minkowski's inequality we obtain

$$\left(E \left\| \Delta \mathbf{W}_i \right\|^4 \right)^{1/2} \leq T + \left(E \left\| \mathbf{Z}_i \right\|^4 \right)^{1/2} + \left(E \left\| \Delta \mathbf{y}_i \right\|^4 \right)^{1/2} .$$

Also $\left\| \mathbf{Z}_i \right\|^2 = \left\| \Delta \mathbf{x}_{i1} \right\|^2 + 2 \sum_{t=2}^{T} \left\| \Delta \mathbf{x}_{it} \right\|^2$, and since by result (16) of Proposition 2 $\sup_{i,t} E\left(\left\| \Delta \mathbf{x}_{it} \right\|^{4+\epsilon} \right) < K$, it then follows that $\sup_i E \left\| \mathbf{Z}_i \right\|^4 < K$. Similarly, using (39), we have

$$\left\| \Delta \mathbf{y}_i \right\| \leq \left\| \mathbf{a} \right\| + \left\| \mathbf{B}^{-1}(\gamma) \right\| \left\| \boldsymbol{\delta} \right\| \left\| \mathbf{Z}_i \right\| + \left\| \mathbf{B}^{-1}(\gamma) \right\| \left\| \boldsymbol{\xi}_i \right\| ,$$

and by assumption $\left\| \mathbf{a} \right\| < K$, $\left\| \boldsymbol{\delta} \right\| < K$, and $\left\| \mathbf{B}^{-1}(\gamma) \right\| < K$. Also by result (15) of Proposition 2 $\sup_{i,t} E \left| \xi_{it} \right|^{4+\epsilon} < K$, and it is already established that $\sup_i E \left\| \mathbf{Z}_i \right\|^4 < K$. Hence,

$$\left(E \left\| \Delta \mathbf{y}_i \right\|^4 \right)^{1/4} \leq \left\| \mathbf{a} \right\| + \left\| \mathbf{B}^{-1}(\gamma) \right\| \left\| \boldsymbol{\delta} \right\| \left(E \left\| \mathbf{Z}_i \right\|^4 \right)^{1/4} + \left\| \mathbf{B}^{-1}(\gamma) \right\| \left(E \left\| \boldsymbol{\xi}_i \right\|^4 \right)^{1/4} ,$$

and it follows that $\sup_i E \left\| \Delta \mathbf{y}_i \right\|^4 < K$, as required. ■

Lemma 2 *Consider the model given by (13) and let*

$$\boldsymbol{\xi}_i(\boldsymbol{\varphi}) = \Delta\mathbf{y}_i - \Delta\mathbf{W}_i\boldsymbol{\varphi},\ \ \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}) = E\left[\boldsymbol{\xi}_i(\boldsymbol{\varphi})\boldsymbol{\xi}_i'(\boldsymbol{\varphi})\right].$$

Define

$$\mathbf{d}_i(\boldsymbol{\psi},\boldsymbol{\varphi}_0) = \Delta\mathbf{W}_i'\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0), \tag{A.4}$$

and suppose that Assumptions 1-6 and Assumption 7(i)-(ii) hold. Then

$$E_0\left[\mathbf{d}_i(\boldsymbol{\psi},\boldsymbol{\varphi}_0)\right] = \mathbf{b}(\boldsymbol{\psi},\boldsymbol{\varphi}_0) = \left[\mathbf{0},\mathbf{0},-\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0)\right]', \tag{A.5}$$

where

$$\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0) = Tr\left\{\left[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}) - \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)\right]\mathbf{C}(\boldsymbol{\psi},\gamma_0)\right\} \tag{A.6}$$

and

$$\mathbf{C}(\boldsymbol{\psi},\gamma_0) = \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_0^{T-3} & \gamma_0^{T-4} & \cdots & 0 & 0 \\ \gamma_0^{T-2} & \gamma_0^{T-3} & \cdots & 1 & 0 \end{pmatrix}. \tag{A.7}$$

Furthermore

$$E_0\left[\mathbf{d}_i(\boldsymbol{\psi}_0,\boldsymbol{\varphi}_0)\right] = \mathbf{0},\ \textit{for}\ i = 1,2,...,N, \tag{A.8}$$

$$\mathbf{b}_N(\boldsymbol{\psi},\boldsymbol{\varphi}_0) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i(\boldsymbol{\psi},\boldsymbol{\varphi}_0) \overset{a.s.}{\to} \mathbf{b}(\boldsymbol{\psi},\boldsymbol{\varphi}_0) = \left[\mathbf{0},\mathbf{0},-\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0)\right]', \tag{A.9}$$

$$\mathbf{b}_N(\boldsymbol{\psi}_0,\boldsymbol{\varphi}_0) = \frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{W}_i'\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0) \overset{a.s.}{\to} \mathbf{0}, \tag{A.10}$$

and

$$\boldsymbol{\Sigma}_{N,\xi}(\boldsymbol{\psi}_0) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)' \overset{a.s.}{\to} \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0). \tag{A.11}$$

Proof. Under (13),

$$\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0) = \Delta\mathbf{y}_i - \Delta\mathbf{W}_i\boldsymbol{\varphi}_0 = \mathbf{G}_0\boldsymbol{\eta}_{0i} + \mathbf{r}_{0i}, \tag{A.12}$$

where $\mathbf{G}_0, \boldsymbol{\eta}_{0i}$, and $\mathbf{r}_{0i}$ denote the values of $\mathbf{G}$, $\boldsymbol{\eta}_i$ and $\mathbf{r}_i$ evaluated at $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. It is now easily seen that $E_0\left[\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)\right] = \mathbf{0}$, and $Var\left[\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)\right] = E_0\left[\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)\boldsymbol{\xi}_i'(\boldsymbol{\varphi}_0)\right] = \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)$. Also under Assumptions 1-6, $\boldsymbol{\xi}_i(\boldsymbol{\varphi}) = \mathbf{G}\boldsymbol{\eta}_i + \mathbf{r}_i$ are independently distributed over i for all values of $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon$, and $\Delta\mathbf{x}_{it}$ is independently distributed from u_{it} and $\boldsymbol{\eta}_i$. Partition $\Delta\mathbf{W}_i$ as $\Delta\mathbf{W}_i = (\mathbf{I}_T, \mathbf{Z}_i, \Delta\widetilde{\mathbf{y}}_{i,-1})$, where $\mathbf{I}_T$ is the identity matrix of order T,

$$\mathbf{Z}_i = \begin{pmatrix} \Delta\mathbf{x}_i' & 0 \\ \mathbf{0} & \Delta\mathbf{x}_{i2}' \\ \vdots & \vdots \\ \mathbf{0} & \Delta\mathbf{x}_{iT}' \end{pmatrix},\ \Delta\widetilde{\mathbf{y}}_{i,-1} = \begin{pmatrix} 0 \\ \Delta y_{i1} \\ \vdots \\ \Delta y_{i,T-1} \end{pmatrix},$$

and note that $\Delta\widetilde{\mathbf{y}}_{i,-1} = \mathbf{L}\Delta\mathbf{y}_i$, where $\mathbf{L}$ is given by (A.3). Also, using (39) and evaluating it at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ we have

$$\Delta\mathbf{y}_i = \mathbf{B}\left(\gamma_0\right)^{-1}\left(\mathbf{Z}_i\boldsymbol{\delta}_0 + \mathbf{d}_0\right) + \mathbf{B}\left(\gamma_0\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right), \tag{A.13}$$

where $\boldsymbol{\delta} = \left(\boldsymbol{\pi}', \boldsymbol{\beta}'\right)'$, and $\mathbf{B}\left(\gamma\right)$ is defined by (35). Consider now (A.4), and note that

$$\mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0) = \boldsymbol{\Delta}\mathbf{W}_i'\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right) = \begin{pmatrix} \boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right) \\ \mathbf{Z}_i'\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right) \\ \Delta\mathbf{y}_i'\mathbf{L}'\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right) \end{pmatrix} = \begin{pmatrix} \mathbf{d}_{1i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right) \\ \mathbf{d}_{2i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right) \\ d_{3i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right) \end{pmatrix}. \tag{A.14}$$

Further, using (A.13), write $d_{3i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right)$ as

$$\begin{aligned} d_{3i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right) &= \left[\mathbf{B}\left(\gamma_0\right)^{-1}\left(\mathbf{Z}_i\boldsymbol{\delta}_0 + \mathbf{d}_0\right) + \mathbf{B}\left(\gamma_0\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right)\right]'\mathbf{L}'\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right) \qquad \text{(A.15)} \\ &= \left(\mathbf{Z}_i\boldsymbol{\delta}_0 + \mathbf{d}_0\right)'\mathbf{B}\left(\gamma_0\right)'^{-1}\mathbf{L}'\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right) + \boldsymbol{\xi}_i'\left(\boldsymbol{\varphi}_0\right)\mathbf{B}\left(\gamma_0\right)'^{-1}\mathbf{L}'\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right). \end{aligned}$$

It is now easily seen that since $E_0\left[\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right)\right] = \mathbf{0}$ and $\mathbf{Z}_i$ is distributed independently from $\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right)$, then

$$E_0\left[\mathbf{d}_{1i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right)\right] = \mathbf{0}, \text{ and } E_0\left[\mathbf{d}_{2i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right)\right] = \mathbf{0}, \text{ for all } i, \tag{A.16}$$

and

$$\begin{aligned} E_0\left[d_{3i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right)\right] &= E_0\left[\boldsymbol{\xi}_i'\left(\boldsymbol{\varphi}_0\right)\mathbf{B}\left(\gamma_0\right)'^{-1}\mathbf{L}'\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right)\right] \\ &= Tr\left\{\mathbf{B}\left(\gamma_0\right)'^{-1}\mathbf{L}'\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}E_0\left[\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right)\boldsymbol{\xi}_i'\left(\boldsymbol{\varphi}_0\right)\right]\right\} \\ &= Tr\left[\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}_0\right)\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\mathbf{L}\mathbf{B}\left(\gamma_0\right)^{-1}\right]. \end{aligned}$$

Also, using (36) and (A.3), we have

$$\mathbf{L}\mathbf{B}\left(\gamma_0\right)^{-1} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_0^{T-3} & \gamma_0^{T-4} & \cdots & 0 & 0 \\ \gamma_0^{T-2} & \gamma_0^{T-3} & \cdots & 1 & 0 \end{pmatrix}.$$

Hence, $Tr\left[\mathbf{L}\mathbf{B}\left(\gamma_0\right)^{-1}\right] = \mathbf{0}$, and $E_0\left[d_{3i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right)\right]$ can be written as

$$E_0\left[d_{3i}\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right)\right] = -Tr\left\{\left[\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right) - \boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}_0\right)\right]\mathbf{C}\left(\boldsymbol{\psi}, \gamma_0\right)\right\} = -\kappa\left(\boldsymbol{\psi}, \boldsymbol{\psi}_0\right), \tag{A.17}$$

where $\mathbf{C}\left(\boldsymbol{\psi}, \gamma_0\right) = \boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\mathbf{L}\mathbf{B}\left(\gamma_0\right)^{-1}$. Using (A.17) and (A.16) now yields (A.5), as required. Result (A.8) then follows immediately, noting that $E_0\left[d_{3i}\left(\boldsymbol{\psi}_0, \boldsymbol{\varphi}_0\right)\right] = Tr\left[\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}_0\right)\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}_0\right)^{-1}\mathbf{L}\mathbf{B}\left(\gamma_0\right)^{-1}\right] = Tr\left[\mathbf{L}\mathbf{B}\left(\gamma_0\right)^{-1}\right] = 0$. To establish (A.9), we first note that $\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right)$, for $i = 1, 2, ..., N$ are independently distributed, and therefore conditional on $\mathbf{Z}_i$, $\mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0)$ are also independently distributed across i. Hence to show that $\mathbf{b}_N\left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0)$ converges almost surely to $\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}E_0\left[\mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0)\right]$, it is sufficient to show that $\sup_i E_0\left\|\mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0)\right\|^2 < K$. Consider each of the three terms of $\mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0)$ in turn. First, from result (A.1) and Liapunov's inequality we have that $E\left\|\xi_i\right\|^2 < K < \infty$ and noting that by assumption 7(ii) $\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}$ is positive definite for all $\boldsymbol{\psi} \in \boldsymbol{\Theta}_\psi$, then

$$\sup_i E_0\left\|\mathbf{d}_{1i}(\boldsymbol{\psi}, \boldsymbol{\varphi}_0)\right\|^2 \leq \left\|\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\right\|^2\sup_i E_0\left\|\boldsymbol{\xi}_i\left(\boldsymbol{\varphi}_0\right)\right\|^2 < K. \tag{A.18}$$

Similarly, using in addition result (A.2) we have

$$\sup_i E_0 \left\| \mathbf{d}_{2i}(\boldsymbol{\psi}, \boldsymbol{\varphi}_0) \right\|^2 \leq \sup_i E \left\| \mathbf{Z}_i \right\|^2 \left\| \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \right\|^2 \sup_i E_0 \left\| \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^2 < K. \tag{A.19}$$

Finally, applying the Minkowski inequality to (A.15) we have

$$\begin{aligned} \left[E_0 \left\| d_{3i}(\boldsymbol{\psi}, \boldsymbol{\varphi}_0) \right\|^2 \right]^{1/2} &\leq \left[E_0 \left\| (\mathbf{Z}_i \boldsymbol{\delta}_0 + \mathbf{d}_0)' \mathbf{B} \left(\gamma_0 \right)'^{-1} \mathbf{L}' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^2 \right]^{1/2} \\ &\quad + \left[E_0 \left\| \boldsymbol{\xi}_i' \left(\boldsymbol{\varphi}_0 \right) \mathbf{B} \left(\gamma_0 \right)'^{-1} \mathbf{L}' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^2 \right]^{1/2}, \end{aligned}$$

$$\begin{aligned} E_0 \left\| (\mathbf{Z}_i \boldsymbol{\delta}_0 + \mathbf{d}_0)' \mathbf{B} \left(\gamma_0 \right)'^{-1} \mathbf{L}' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^2 &\leq E_0 \left\| \mathbf{Z}_i \boldsymbol{\delta}_0 + \mathbf{d}_0 \right\|^2 \left\| \mathbf{B} \left(\gamma_0 \right)'^{-1} \mathbf{L}' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \right\|^2 \\ &\quad \times E_0 \left\| \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^2, \\ E_0 \left\| \boldsymbol{\xi}_i' \left(\boldsymbol{\varphi}_0 \right) \mathbf{B} \left(\gamma_0 \right)'^{-1} \mathbf{L}' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi}_0 \right)^{-1} \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^2 &\leq \left\| \mathbf{B} \left(\gamma_0 \right)'^{-1} \mathbf{L}' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \right\|^2 E_0 \left\| \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^4. \end{aligned}$$

But $\left\| \mathbf{B} \left(\gamma_0 \right)'^{-1} \mathbf{L}' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \right\|^2 \leq \left\| \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \right\|^2 \left\| \mathbf{L} \right\|^2 \left\| \mathbf{B} \left(\gamma_0 \right)^{-1} \right\|^2$, and it is easily seen that $\left\| \mathbf{L} \right\| \leq 1$, and $\left\| \mathbf{B} \left(\gamma_0 \right)^{-1} \right\| \leq \sum_{t=1}^{T} \left| \gamma_0 \right|^{t-1} < K$. Also, by results of Lemma 1, $\sup_i E_0 \left\| \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^4 < K$, and $\left\| \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \right\| < K$, by assumption. Further, $E_0 \left\| \mathbf{Z}_i \boldsymbol{\delta}_0 + \mathbf{d}_0 \right\|^2 \leq \left\| \boldsymbol{\delta}_0 \right\|^2 E \left\| \mathbf{Z}_i \right\|^2 + \left\| \mathbf{d}_0 \right\|^2$ which is uniformly bounded under results (A.2) of Lemma 1, noting that $\boldsymbol{\delta}_0$ and $\mathbf{d}_0$ are defined on a compact set and are bounded as well. Therefore, $\sup_i E_0 \left\| d_{3i}(\boldsymbol{\psi}, \boldsymbol{\varphi}_0) \right\|^2 < K$. Now using this result together with (A.18) and (A.19) in (A.14) we have

$$\sup_i E_0 \left\| \mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0) \right\|^2 = \sup_i E_0 \left\| \Delta \mathbf{W}_i' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \right\|^2 < K,$$

which establishes that $\mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0)$ is uniformly L_2-bounded, besides being cross-sectionally independent. Hence,

$$\mathbf{b}_N \left(\boldsymbol{\psi}, \boldsymbol{\varphi}_0 \right) = N^{-1} \sum_{i=1}^{N} \mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0) \overset{a.s.}{\rightarrow} \lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} E_0 \left[\mathbf{d}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_0) \right] = \left[\mathbf{0}, \mathbf{0}, -\kappa \left(\boldsymbol{\psi}, \boldsymbol{\psi}_0 \right) \right]',$$

which establishes (A.9). Result (A.10) follows from the above by setting $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ and noting from (A.8) that $E_0 \left[\mathbf{d}_i(\boldsymbol{\psi}_0, \boldsymbol{\varphi}_0) \right] = \mathbf{0}$. Finally, since $\sup_i E_0 \left\| \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \boldsymbol{\xi}_i' \left(\boldsymbol{\varphi}_0 \right) \right\|^2 < K$, for a finite T (see result (A.1) of Lemma 1), and by assumption $\boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \boldsymbol{\xi}_i' \left(\boldsymbol{\varphi}_0 \right)$ are distributed independently over i, then

$$\boldsymbol{\Sigma}_{N,\xi} \left(\boldsymbol{\psi}_0 \right) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right)' \overset{a.s.}{\rightarrow} \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E_0 \left[\boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right)' \right],$$

and result (A.11) follows, since $E_0 \left[\boldsymbol{\xi}_i \left(\boldsymbol{\varphi}_0 \right) \boldsymbol{\xi}_i' \left(\boldsymbol{\varphi}_0 \right) \right] = \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi}_0 \right)$. ■

Lemma 3 *Consider the average log-likelihood function*

$$\bar{\ell}_N \left(\boldsymbol{\theta} \right) = \bar{\ell}_N \left(\boldsymbol{\varphi}, \boldsymbol{\psi} \right) = -\frac{T}{2} \ln \left(2\pi \right) - \frac{1}{2} \ln \left| \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right) \right| - \frac{1}{2N} \sum_{i=1}^{N} \boldsymbol{\xi}_i \left(\boldsymbol{\varphi} \right)' \boldsymbol{\Sigma}_\xi \left(\boldsymbol{\psi} \right)^{-1} \boldsymbol{\xi}_i \left(\boldsymbol{\varphi} \right) \tag{A.20}$$

$\bar{\ell}_N(\boldsymbol{\theta}) = N^{-1}\ell_N(\boldsymbol{\theta})$ *and* $\ell_N(\boldsymbol{\theta})$ *is defined by (20). Then under Assumptions 1-7 we have*

$$\bar{\ell}_N(\boldsymbol{\theta}_0) \overset{a.s.}{\to} -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)| - \frac{T}{2}, \tag{A.21}$$

and

$$\begin{aligned}\bar{\ell}_N(\boldsymbol{\theta}) &\overset{a.s.}{\to} -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| - \frac{1}{2}Tr\left[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)\right] \\ &-\frac{1}{2}(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)'\mathbf{A}(\boldsymbol{\psi})(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0) - (\gamma-\gamma_0)\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0),\end{aligned} \tag{A.22}$$

where $\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0)$ *is defined by (A.6). Also*

$$\bar{\ell}_N(\boldsymbol{\theta}_0) - \bar{\ell}_N(\boldsymbol{\theta}) \overset{a.s.}{\to} \lim_{N\to\infty} E_0\left[\bar{\ell}_N(\boldsymbol{\theta}_0) - \bar{\ell}_N(\boldsymbol{\theta})\right] \geq 0, \tag{A.23}$$

where

$$\begin{aligned}\lim_{N\to\infty} E_0\left[\bar{\ell}_N(\boldsymbol{\theta}_0) - \bar{\ell}_N(\boldsymbol{\theta})\right] &= \frac{1}{2}Tr\left[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)\right] - \frac{1}{2}\log\left(|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)| / |\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})|\right) - \frac{T}{2} \\ &+\frac{1}{2}(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)'\mathbf{A}(\boldsymbol{\psi})(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0) + (\gamma-\gamma_0)\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0).\end{aligned} \tag{A.24}$$

Proof. Result (A.21) follows by evaluating (A.20) under $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and using (A.11) from Lemma 2. To establish (A.22) we first note that for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon$, $\boldsymbol{\xi}_i(\boldsymbol{\varphi}) = \boldsymbol{\xi}_i(\boldsymbol{\varphi}_0) - \Delta\mathbf{W}_i(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)$, and using this result in (A.20) we have

$$\begin{aligned}\bar{\ell}_N(\boldsymbol{\theta}) &= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| - \frac{1}{2N}\left[\begin{array}{c}\sum_{i=1}^N\left[\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0) - \Delta\mathbf{W}_i(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)\right]'\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1} \\ \times\left[\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0) - \Delta\mathbf{W}_i(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)\right]\end{array}\right] \\ &= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| - \frac{1}{2}\left[\begin{array}{c}Tr\left(\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\left[\frac{1}{N}\sum_{i=1}^N\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)'\right]\right) \\ -2(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)'\mathbf{b}_N(\boldsymbol{\psi},\boldsymbol{\varphi}_0) \\ +(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)'\mathbf{A}_N(\boldsymbol{\psi})(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)\end{array}\right],\end{aligned} \tag{A.25}$$

where

$$\mathbf{A}_N(\boldsymbol{\psi}) = \frac{1}{N}\sum_{i=1}^N\Delta\mathbf{W}_i'\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\Delta\mathbf{W}_i,\ \mathbf{b}_N(\boldsymbol{\psi},\boldsymbol{\varphi}_0) = \frac{1}{N}\sum_{i=1}^N\mathbf{d}_i(\boldsymbol{\psi},\boldsymbol{\varphi}_0), \tag{A.26}$$

and $\mathbf{d}_i(\boldsymbol{\psi},\boldsymbol{\varphi}_0) = \Delta\mathbf{W}_i'\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)$, as defined by (A.4). Result (A.22) follows using (A.9) and (A.11) from Lemma 2 in (A.25) evaluated at $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$, respectively. Results (A.23) and (A.24) follow from the sure convergence property of (A.21) and (A.22), and the Kullback–Leibler type information inequality. ■

Lemma 4 *Consider the average log-likelihood function defined by (41) and (22):*

$$\begin{aligned}\bar{\ell}_N(\boldsymbol{\theta}) &= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| - \frac{1}{2N}\sum_{i=1}^N\boldsymbol{\xi}_i'(\boldsymbol{\varphi})\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}), \\ \boldsymbol{\xi}_i(\boldsymbol{\varphi}) &= \Delta\mathbf{y}_i - \Delta\mathbf{W}_i\boldsymbol{\varphi},\end{aligned}$$

and suppose that Assumptions 1 to 6 and Assumption 7(i)-(ii) hold. Denote the average score function by $\bar{\mathbf{s}}_N(\boldsymbol{\theta}) = \partial\bar{\ell}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$*. Then*

$$\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0) \overset{a.s}{\to} \mathbf{0}, \tag{A.27}$$

$$\sqrt{N}\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0) \to_d N\left[\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)\right], \tag{A.28}$$

where

$$\mathbf{J}(\boldsymbol{\theta}_0) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} E\left[\boldsymbol{\omega}_i(\boldsymbol{\theta}_0)\,\boldsymbol{\omega}_i'(\boldsymbol{\theta}_0)\right], \tag{A.29}$$

$$\boldsymbol{\omega}_i(\boldsymbol{\theta}_0) = \begin{pmatrix} \Delta\mathbf{W}_i'\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0) \\ \boldsymbol{\nu}_i(\boldsymbol{\theta}_0) \end{pmatrix}, \tag{A.30}$$

with the j^{th} *element of* $\boldsymbol{\nu}_i(\boldsymbol{\theta}_0)$ *given by*

$$\nu_{ij}(\boldsymbol{\theta}_0) = \frac{1}{2}\boldsymbol{\xi}_i'(\boldsymbol{\varphi}_0)\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\frac{\partial\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial\psi_j}\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0) - \frac{1}{2}Tr\left[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\frac{\partial\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial\psi_j}\right]. \tag{A.31}$$

A consistent estimator of $\mathbf{J}(\boldsymbol{\theta}_0)$ *is given by*

$$\hat{\mathbf{J}}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\omega}_i\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{\omega}_i'\left(\hat{\boldsymbol{\theta}}\right), \tag{A.32}$$

where $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon}\bar{\ell}_N(\boldsymbol{\theta})$*.*

Proof. Let $\bar{\mathbf{s}}_N(\boldsymbol{\theta}) = \left(\bar{\mathbf{s}}_{N,\varphi}'(\boldsymbol{\theta}), \bar{\mathbf{s}}_{N,\psi}'(\boldsymbol{\theta})\right)'$, $\boldsymbol{\psi} = \left(\psi_1, \psi_2, ..., \psi_{n_\psi}\right)'$, where $n_\psi = \dim(\boldsymbol{\psi}) = 1 + Tm - m(m-1)/2$, and note that

$$\begin{aligned}
\bar{\mathbf{s}}_{N,\varphi}(\boldsymbol{\theta}) &= \frac{\partial\bar{\ell}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\varphi}} = \frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{W}_i'\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}), \\
\bar{\mathbf{s}}_{N,\psi_j}(\boldsymbol{\theta}) &= \frac{\partial\bar{\ell}_N(\boldsymbol{\theta})}{\partial\psi_j} = -\frac{1}{2}\frac{\partial\ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})|}{\partial\psi_j} + \frac{1}{2N}\sum_{i=1}^{N}\boldsymbol{\xi}_i'(\boldsymbol{\varphi})\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\frac{\partial\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})}{\partial\psi_j}\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}),
\end{aligned}$$

for $j = 1, 2, ..., n_\psi$. Using (A.4), and result (A.10) of Lemma 2, it then readily follows that

$$\bar{\mathbf{s}}_{N,\varphi}(\boldsymbol{\theta}_0) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i(\boldsymbol{\theta}_0) \overset{a.s}{\to} \mathbf{0}, \tag{A.33}$$

Also

$$E_0\left[\boldsymbol{\xi}_i'(\boldsymbol{\varphi}_0)\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\frac{\partial\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial\psi_j}\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)\right] = Tr\left[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\frac{\partial\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial\psi_j}\right],$$

and using well know results on the partial derivatives of the determinants we have (see, for example, Magnus and Neudecker (1988, p.151)).

$$\frac{\partial\ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)|}{\partial\psi_j} = Tr\left[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\frac{\partial\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial\psi_j}\right],$$

and hence $\bar{\mathbf{s}}_{N,\psi}(\boldsymbol{\theta})$ can be written alternatively as

$$\bar{\mathbf{s}}_{N,\psi_j}(\boldsymbol{\theta}_0) = \frac{\partial \bar{\ell}_N(\boldsymbol{\theta}_0)}{\partial \psi_j} = \frac{1}{N}\sum_{i=1}^{N} \nu_{ij}.$$

where

$$\nu_{ij}(\boldsymbol{\theta}_0) = \frac{1}{2}\boldsymbol{\xi}_i'(\boldsymbol{\varphi}_0)\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\frac{\partial \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial \psi_j}\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0) - \frac{1}{2}Tr\left[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\frac{\partial \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial \psi_j}\right]. \tag{A.34}$$

Therefore,

$$\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0) = \begin{pmatrix} \bar{\mathbf{s}}_{N,\varphi}(\boldsymbol{\theta}_0) \\ \bar{\mathbf{s}}_{N,\psi}(\boldsymbol{\theta}_0) \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i(\boldsymbol{\theta}_0) \\ \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\nu}_i(\boldsymbol{\theta}_0) \end{pmatrix},$$

where $\boldsymbol{\nu}_i(\boldsymbol{\theta}_0) = \left(\nu_{i1}(\boldsymbol{\theta}_0), \nu_{i2}(\boldsymbol{\theta}_0), ..., \nu_{i,n_\psi}(\boldsymbol{\theta}_0)\right)'$.

$$\sup_i E\left\|\boldsymbol{\nu}_i(\boldsymbol{\theta}_0)\right\|^2 = \sup_i E\left(\boldsymbol{\nu}_i'(\boldsymbol{\theta}_0)\boldsymbol{\nu}_i(\boldsymbol{\theta}_0)\right) = \sum_{j=1}^{n_\psi}\sup_i E\left(\nu_{ij}^2(\boldsymbol{\theta}_0)\right) \le n_\psi \sup_{i,j} E\left|\nu_{ij}(\boldsymbol{\theta}_0)\right|^2,$$

and application of Minkowski's inequality to (A.34) yields

$$\sup_i E\left|\nu_{ij}(\boldsymbol{\theta}_0)\right|^2 \le \frac{1}{4}\left[\left\|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\right\|^2\left\|\frac{\partial \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial \psi_j}\right\|\left(\sup_i E\left\|\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)\right\|^4\right)^{1/2} + |C|\right]^2,$$

where $C = Tr\left[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\frac{\partial \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial \psi_j}\right]$. But under Assumption 7(ii) and noting that n_ψ is finite, we also have $\left\|\frac{\partial \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)}{\partial \psi_j}\right\| < K$, and $\left\|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi}_0)^{-1}\right\| < K$, and from result (A.1) $\sup_i E\left\|\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)\right\|^4 < K$. Therefore, $\sup_i E\left\|\boldsymbol{\nu}_i(\boldsymbol{\theta}_0)\right\|^2 < K$. Also recall that $\boldsymbol{\xi}_i(\boldsymbol{\varphi}_0)$ are independently distributed over i, which implies that $\boldsymbol{\nu}_i$ are also independently distributed across i. Therefore, $\boldsymbol{\nu}_i$ have zero means (by construction), are independently distributed over i and have bounded second-order moments, which ensure that $\bar{\mathbf{s}}_{N,\psi}(\boldsymbol{\theta}_0) \overset{a.s}{\to} \mathbf{0}$, and together with (A.33) yields $\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0) \overset{a.s}{\to} \mathbf{0}$, as required. Consider now the limiting distribution of $\sqrt{N}\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0)$ and note that

$$\sqrt{N}\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0) = \begin{pmatrix} \sqrt{N}\bar{\mathbf{s}}_{N,\varphi}(\boldsymbol{\theta}_0) \\ \sqrt{N}\bar{\mathbf{s}}_{N,\psi}(\boldsymbol{\theta}_0) \end{pmatrix} = \frac{1}{\sqrt{N}}\begin{pmatrix} \sum_{i=1}^{N}\mathbf{d}_i(\boldsymbol{\theta}_0) \\ \sum_{i=1}^{N}\boldsymbol{\nu}_i(\boldsymbol{\theta}_0) \end{pmatrix} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\omega}_i(\boldsymbol{\theta}_0),$$

where $\boldsymbol{\omega}_i(\boldsymbol{\theta}_0) = \left(\mathbf{d}_i'(\boldsymbol{\theta}_0), \boldsymbol{\nu}_i'(\boldsymbol{\theta}_0)\right)'$, and it is already established that $\boldsymbol{\omega}_i(\boldsymbol{\theta}_0)$ are independently distributed over i, have zero means and bounded second-order moments. Therefore, by the Liapounov central limit theorem and the Cramér-Wold device we have[6] $\sqrt{N}\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0) \to_d N\left[\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)\right]$,where $\mathbf{J}(\boldsymbol{\theta}_0)$ is given by (A.29), as required. Consistency of $\hat{\mathbf{J}}\left(\hat{\boldsymbol{\theta}}\right)$ for $\mathbf{J}(\boldsymbol{\theta}_0)$ follows from consistency of $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$, and the independence of $\boldsymbol{\omega}_i(\boldsymbol{\theta}_0)$ over i. ■

[6] See, for example, Theorem 5.10 in White (1984).

A.2 Proofs of Propositions and Theorems

Proof of Proposition 1. Using (5), and starting from some arbitrary initial state, $\Delta y_{i,-S+1}$, we obtain the following expression

$$\Delta y_{i1} = \gamma^S \Delta y_{i,-S+1} + \sum_{j=0}^{S-1} \gamma^j \boldsymbol{\beta}' \Delta \mathbf{x}_{i,1-j} + \tilde{d}_1 + \tilde{\mathbf{g}}_1' \boldsymbol{\eta}_i + \sum_{j=0}^{S-1} \gamma^j \Delta u_{i,1-j}, \tag{A.35}$$

where $\tilde{d}_1 = \sum_{j=0}^{S-1} \gamma^j d_{1-j}$, and $\tilde{\mathbf{g}}_1 = \sum_{j=0}^{S-1} \gamma^j \mathbf{g}_{1-j}$. Under Assumptions 3 and 6 and following the procedure in Hsiao et al. (2012), we have

$$E\left(\gamma^S \Delta y_{i,-S+1} + \sum_{j=0}^{S-1} \gamma^j \boldsymbol{\beta}' \Delta \mathbf{x}_{i,1-j} \,|\Delta \mathbf{x}_i \right) = a + \boldsymbol{\pi}' \Delta \mathbf{x}_i, \tag{A.36}$$

where the fixed parameters a and $\boldsymbol{\pi}$ are functions of a_S, $\boldsymbol{\pi}_S$, γ and $\boldsymbol{\beta}$. Let

$$\chi_i = \gamma^S \Delta y_{i,-S+1} + \sum_{j=0}^{S-1} \gamma^j \boldsymbol{\beta}' \Delta \mathbf{x}_{i,1-j} - E\left(\gamma^S \Delta y_{i,-S+1} + \sum_{j=0}^{S-1} \gamma^j \boldsymbol{\beta}' \Delta \mathbf{x}_{i,1-j} \,|\Delta \mathbf{x}_i \right), \tag{A.37}$$

where by construction χ_i is a martingale difference process. Also in view of Assumptions 3 and 6 and by application of the Minkowski inequality to both sides of χ_i we have $\sup_i |\chi_i|^{4+\epsilon} < K$.[7] Hence, using (A.36) and (A.37) in (A.35) we have

$$\Delta y_{i1} = d_1 + \boldsymbol{\pi}' \Delta \mathbf{x}_i + \tilde{\mathbf{g}}_1' \boldsymbol{\eta}_i + v_{i1},$$

where $d_1 = a + \tilde{d}_1$, and

$$v_{i1} = \sum_{j=0}^{S-1} \gamma^j \Delta u_{i,1-j} + \chi_i. \tag{A.38}$$

It now readily follows that $v_{i1} \sim IID(0, \omega\sigma^2)$, where $\omega\sigma^2 = Var\left(v_{i1}\right)$, and v_{i1} is distributed independently of $\Delta \mathbf{x}_i$ and $\boldsymbol{\eta}_i$. Again by application of the Minkowski inequality to (A.38) we also have that $\sup_i |v_{i1}|^{4+\epsilon} < K$, as required. Further, under Assumptions 3 and 6, $\sup_i Var\left(\chi_i\right) < K$ and as a result $0 < \omega_{\min} < \omega < \omega_{\max} < \infty$, where $\omega_{\min}$ and $\omega_{\max}$ are fixed constants. Finally, it is easily established that

$$\begin{aligned} Cov\left(v_{i1}, \Delta u_{it}\right) &= -\sigma^2, \text{ for } t = 2 \\ &= 0, \text{ for } t = 3, 4, ..., T, \end{aligned}$$

as required. ■

Proof of Proposition 2. Result (15) follows by applying Minkowski's inequality to the elements of $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, \cdots, \xi_{iT})'$. Specifically, for $t = 2, 3, ..., T$, $\xi_{it} = \mathbf{g}_t' \boldsymbol{\eta}_i + \Delta u_{it}$ and we

[7] Note that under Assumption 3 $\sup_{i,t} E \left\| \Delta \mathbf{x}_{it} \right\|^{4+\epsilon} < K$, which also follows from application of the Minkowski inequality for infinite sums to (3).

have

$$\begin{aligned}\left(E\left|\xi_{it}\right|^{4+\epsilon}\right)^{\frac{1}{4+\epsilon}} &= \left(E\left|\mathbf{g}_t'\boldsymbol{\eta}_i+\Delta u_{it}\right|^{4+\epsilon}\right)^{\frac{1}{4+\epsilon}}\\ &\leq \left(E\left|\mathbf{g}_t'\boldsymbol{\eta}_i\right|^{4+\epsilon}\right)^{\frac{1}{4+\epsilon}}+\left(E\left|\Delta u_{it}\right|^{4+\epsilon}\right)^{\frac{1}{4+\epsilon}}\\ &\leq \left\|\mathbf{g}_t\right\|\left(E\left\|\boldsymbol{\eta}_i\right\|^{4+\epsilon}\right)^{\frac{1}{4+\epsilon}}+\left(E\left|\Delta u_{it}\right|^{4+\epsilon}\right)^{\frac{1}{4+\epsilon}}.\end{aligned}$$

Under Assumptions 1, 2 and 5 $\sup_t\left\|\mathbf{g}_t\right\|<K$, $\sup_i E\left\|\boldsymbol{\eta}_i\right\|^{4+\epsilon}<K$ and $\sup_{i,t}E\left|\Delta u_{it}\right|^{4+\epsilon}$. Similarly for $t=1$, where $\xi_{i1}=\tilde{\mathbf{g}}_1'\boldsymbol{\eta}_i+v_{i1}$ and for which it holds under the assumptions of Proposition 1 that $\left\|\mathbf{g}_1\right\|<K$ and $\sup_i E\left|v_{i1}\right|^{4+\epsilon}<K$. Hence, $\left(E\left|\xi_{it}\right|^{4+\epsilon}\right)^{\frac{1}{4+\epsilon}}\leq K$, for $t=1,2,...,T$ and (15) follows as required. To establish condition (16), using (3) we first note that

$$\left\|\Delta\mathbf{x}_{it}\right\|\leq\left\|\mathbf{c}_x\right\|+\sum_{j=0}^{\infty}\left\|\boldsymbol{\Psi}_j\right\|\left\|\boldsymbol{\varepsilon}_{i,t-j}\right\|,$$

and by the Minkowski inequality for infinite sums we have

$$\left(E\left\|\Delta\mathbf{x}_{it}\right\|^p\right)^{1/p}\leq\left\|\mathbf{c}_x\right\|+\sum_{j=0}^{\infty}\left\|\boldsymbol{\Psi}_j\right\|\left(E\left\|\boldsymbol{\varepsilon}_{i,t-j}\right\|^p\right)^{1/p},$$

for any $p\geq 1$. Set $p=4+\epsilon$, and note that under Assumption 3, $\left\|\mathbf{c}_x\right\|<K$, $\sup_{i,t}E\left\|\boldsymbol{\varepsilon}_{it}\right\|^{4+\epsilon}<K$, and $\sum_{j=0}^{\infty}\left\|\boldsymbol{\Psi}_j\right\|<K$. Therefore, $\left(E\left\|\Delta\mathbf{x}_{it}\right\|^{4+\epsilon}\right)^{1/(4+\epsilon)}\leq K$, and (16) follows as required.

■

Proof of Proposition 3.

Recall that $\boldsymbol{\theta}=(\boldsymbol{\varphi}',\boldsymbol{\psi}')'$, and $\boldsymbol{\varphi}=(\boldsymbol{\lambda}',\gamma)'$, and using (41) note that

$$\bar{\ell}_N\left(\boldsymbol{\lambda},\gamma,\boldsymbol{\psi}\right)=-\frac{T}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left|\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)\right|-\frac{1}{2N}\sum_{i=1}^{N}\boldsymbol{\xi}_i'(\boldsymbol{\lambda},\gamma)\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i(\boldsymbol{\lambda},\gamma).$$

Using results (A.23) and (A.24)) in Lemma 3 we have

$$\bar{\ell}_N\left(\boldsymbol{\lambda}_0,\gamma_0,\boldsymbol{\psi}_0\right)-\bar{\ell}_N\left(\boldsymbol{\lambda},\gamma,\boldsymbol{\psi}\right)\overset{a.s.}{\to}\lim_{N\to\infty}E_0\left[\bar{\ell}_N\left(\boldsymbol{\lambda}_0,\gamma_0,\boldsymbol{\psi}_0\right)-\bar{\ell}_N\left(\boldsymbol{\lambda},\gamma,\boldsymbol{\psi}\right)\right]\geq 0,\tag{A.39}$$

$$2\lim_{N\to\infty}E_0\left[\bar{\ell}_N\left(\boldsymbol{\lambda}_0,\gamma_0,\boldsymbol{\psi}_0\right)-\bar{\ell}_N\left(\boldsymbol{\lambda},\gamma,\boldsymbol{\psi}\right)\right]=\chi\left(\boldsymbol{\psi},\boldsymbol{\psi}_0\right)+\left(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0\right)'\mathbf{A}\left(\boldsymbol{\psi}\right)\left(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0\right)+2\left(\gamma-\gamma_0\right)\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0),\tag{A.40}$$

where

$$\chi\left(\boldsymbol{\psi},\boldsymbol{\psi}_0\right)=Tr\left[\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}_0\right)\right]-\ln\left(\left|\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}_0\right)\right|/\left|\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)\right|\right)-T,\tag{A.41}$$

and

$$\kappa\left(\boldsymbol{\psi},\boldsymbol{\psi}_0\right)=Tr\left\{\left[\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)-\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}_0\right)\right]\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\mathbf{LB}\left(\gamma_0\right)^{-1}\right\},\tag{A.42}$$

with $\mathbf{B}\left(\gamma_0\right)^{-1}$ given by (36) evaluated at γ_0, $\mathbf{L}$ is a matrix lag operator defined by (A.3) and $\mathbf{A}\left(\boldsymbol{\psi}\right)$ is defined by (42). Denote the eigenvalues of $\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}_0\right)$ and $\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)$ by λ_{0t} and λ_t $(t=1,2,...,T)$, respectively (note that $\lambda_{0t}>0$ and $\lambda_t>0$) and write $\chi\left(\boldsymbol{\psi},\boldsymbol{\psi}_0\right)$ as

$$\chi\left(\boldsymbol{\psi},\boldsymbol{\psi}_0\right)=\sum_{t=1}^{T}\left[\left(\lambda_{0t}/\lambda_t\right)-\ln\left(\lambda_{0t}/\lambda_t\right)-1\right].$$

Also note that $(\lambda_{0t}/\lambda_t) - \ln(\lambda_{0t}/\lambda_t) - 1 \geq 0$ with the equality holding if and only if $\lambda_{0t} = \lambda_t$, for all t, or equivalently if and only if $\boldsymbol{\psi} = \boldsymbol{\psi}_0$.[8] Therefore, $\chi(\boldsymbol{\psi}, \boldsymbol{\psi}_0) \geq 0$, with equality holding if and only if $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Furthermore, since by Assumption 7 (iii) $\mathbf{A}(\boldsymbol{\psi})$ is a positive definite matrix, then

$$(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0)' \mathbf{A}(\boldsymbol{\psi})(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0) \geq \lambda_{\min}[\mathbf{A}(\boldsymbol{\psi})](\boldsymbol{\varphi} - \boldsymbol{\varphi}_0)'(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0),$$

where $\lambda_{\min}[\mathbf{A}(\boldsymbol{\psi})] > 0$. It is clear that the first two terms of (A.40) can not be negative, but the same is not true of the third term, $(\gamma - \gamma_0)\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0)$, and therefore, global identification of γ_0 can not be guaranteed. Consider now the almost sure probability limit of $\bar{\ell}_N(\boldsymbol{\varphi}_0, \boldsymbol{\psi}_0) - \bar{\ell}_N(\boldsymbol{\varphi}, \boldsymbol{\psi})$ on the set $\boldsymbol{\Theta}_\epsilon = \mathcal{N}_\epsilon(\gamma_0) \times \boldsymbol{\Theta}_\varphi \times \boldsymbol{\Theta}_\psi$, for some small positive ϵ, where $\mathcal{N}_\epsilon(\gamma_0)$ is defined by 1. We now establish that there exists $\epsilon > 0$ for which this limit can be zero if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. To see this consider the first and the third terms of (A.40) together, and note that $\chi(\boldsymbol{\psi}, \boldsymbol{\psi}_0) + 2(\gamma - \gamma_0)\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0) = 0$ if $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. In such a case

$$2 \lim_{N\to\infty} E_0\left[\bar{\ell}_N(\boldsymbol{\varphi}_0, \boldsymbol{\psi}_0) - \bar{\ell}_N(\boldsymbol{\varphi}, \boldsymbol{\psi}_0)\right] \geq \frac{1}{2}\lambda_{\min}[\mathbf{A}(\boldsymbol{\psi}_0)](\boldsymbol{\varphi} - \boldsymbol{\varphi}_0)'(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0),$$

and $\bar{\ell}_N(\boldsymbol{\varphi}_0, \boldsymbol{\psi}_0) - \bar{\ell}_N(\boldsymbol{\varphi}, \boldsymbol{\psi}_0) \overset{a.s.}{\to} 0$, if and only if $\lambda_{\min}[\mathbf{A}(\boldsymbol{\psi}_0)](\boldsymbol{\varphi} - \boldsymbol{\varphi}_0)'(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0) = 0$, which implies $\boldsymbol{\varphi} = \boldsymbol{\varphi}_0$, as required since by assumption $\lambda_{\min}[\mathbf{A}(\boldsymbol{\psi}_0)] > 0$. Consider now the case where $\boldsymbol{\psi} \neq \boldsymbol{\psi}_0$, and note that $\chi(\boldsymbol{\psi}, \boldsymbol{\psi}_0) > 0$, and $|\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0)| > 0$, and therefore on $\boldsymbol{\Theta}_\epsilon$ we have

$$|(\gamma - \gamma_0)\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0)| \leq |(\gamma - \gamma_0)| |\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0)| < \epsilon |\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0)|.$$

Also note that under Assumptions 1, 2 and 5, $\|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})\| < K$ for all $\boldsymbol{\psi} \in \boldsymbol{\Theta}_\psi$, and it is readily seen that $|\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0)| < K$. Hence, on $\boldsymbol{\Theta}_\epsilon$ there must exist $\epsilon > 0$, such that $\chi(\boldsymbol{\psi}, \boldsymbol{\psi}_0) + 2(\gamma - \gamma_0)\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0) \geq 0$, and hence

$$2 \lim_{N\to\infty} E_0\left[\bar{\ell}_N(\boldsymbol{\varphi}_0, \boldsymbol{\psi}_0) - \bar{\ell}_N(\boldsymbol{\varphi}, \boldsymbol{\psi})\right] \geq \frac{1}{2}\lambda_{\min}[\mathbf{A}(\boldsymbol{\psi})](\boldsymbol{\varphi} - \boldsymbol{\varphi}_0)'(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0).$$

Once again since by assumption $\lambda_{\min}[\mathbf{A}(\boldsymbol{\psi})] > 0$ for all values of $\boldsymbol{\psi} \in \boldsymbol{\Theta}_\psi$, then on $\boldsymbol{\Theta}_\epsilon$ there exists $\epsilon > 0$ such that $\boldsymbol{\varphi} = \boldsymbol{\varphi}_0$, and hence $\boldsymbol{\psi} = \boldsymbol{\psi}_0$, almost surely. ■

Proof of Theorem 1. For the proof of consistency it suffices to show here that under the assumptions of the theorem, $\bar{C}_N(\boldsymbol{\theta}) = -2\bar{\ell}_N(\boldsymbol{\theta}) \overset{a.s.}{\to} \bar{C}(\boldsymbol{\theta})$ uniformly on $\boldsymbol{\Theta}_\epsilon$ (see Section 6). From results in Lemma 3 (see (A.23) and (A.24)) it follows that $\bar{C}_N(\boldsymbol{\theta}) = -2\bar{\ell}_N(\boldsymbol{\theta}) \overset{a.s.}{\to} \bar{C}(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon$, where

$$\bar{C}_N(\boldsymbol{\theta}) = \bar{C}_N(\boldsymbol{\varphi}, \boldsymbol{\psi}) = T\ln(2\pi) + \ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\xi}_i(\boldsymbol{\varphi})' \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1} \boldsymbol{\xi}_i(\boldsymbol{\varphi})$$

and

$$\bar{C}(\boldsymbol{\theta}) = \bar{C}(\boldsymbol{\varphi}, \boldsymbol{\psi}) = \chi(\boldsymbol{\psi}, \boldsymbol{\psi}_0) + (\boldsymbol{\varphi} - \boldsymbol{\varphi}_0)' \mathbf{A}(\boldsymbol{\psi})(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0) + 2(\gamma - \gamma_0)\kappa(\boldsymbol{\psi}, \boldsymbol{\psi}_0) + C(\boldsymbol{\psi}_0),$$

[8] Note that for any $x > 0$, $\ln(x) \leq x - 1$. Here $x = \lambda_{0t}/\lambda_t > 0$.

and the term $C(\boldsymbol{\psi}_0)$ does not depend on $\boldsymbol{\theta}$. Since under our assumptions $\bar{\ell}_N(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, this pointwise result holds uniformly on $\boldsymbol{\Theta}_\epsilon$ by the uniform law of large numbers, so long as the dominance condition

$$E \sup_{\theta \in \boldsymbol{\Theta}_\epsilon} \left| \boldsymbol{\xi}_i'(\boldsymbol{\varphi}) \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1} \boldsymbol{\xi}_i(\boldsymbol{\varphi}) + T \ln(2\pi) + \ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| \right| < \infty$$

holds; see for example Theorem 23 of Pötscher and Prucha (2001). Since T is finite, it is sufficient to show that

$$E \sup_{\theta \in \boldsymbol{\Theta}_\epsilon} \left| \boldsymbol{\xi}_i'(\boldsymbol{\varphi}) \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1} \boldsymbol{\xi}_i(\boldsymbol{\varphi}) + \ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| \right| < \infty.$$

We have

$$E \sup_{\theta \in \boldsymbol{\Theta}_\epsilon} \left| \boldsymbol{\xi}_i'(\boldsymbol{\varphi}) \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1} \boldsymbol{\xi}_i(\boldsymbol{\varphi}) + \ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| \right| \leq E \sup_{\theta \in \boldsymbol{\Theta}_\epsilon} \left| \boldsymbol{\xi}_i'(\boldsymbol{\varphi}) \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1} \boldsymbol{\xi}_i(\boldsymbol{\varphi}) \right| + \sup_{\psi \in \boldsymbol{\Theta}_\psi} \left| \ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| \right|.$$

Starting with the second term and using Assumption 7(ii) and the property that for any positive definite real $n \times n$ matrix $\mathbf{A}$, $\ln|\mathbf{A}| \leq Tr(\mathbf{A}) - n$,

$$\begin{aligned}
\sup_{\psi \in \boldsymbol{\Theta}_\psi} \left| \ln|\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})| \right| &\leq \sup_{\psi \in \boldsymbol{\Theta}_\psi} Tr[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})] - T \\
&\leq \sup_{\psi \in \boldsymbol{\Theta}_\psi} \left(\textstyle\sum_{t=1}^T \lambda_t[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})] \right) - T \\
&\leq T \sup_{\psi \in \boldsymbol{\Theta}_\psi} (\lambda_{\max}[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})]) - T \\
&\leq T(c_{\max} - 1) < \infty.
\end{aligned}$$

For the first term, defining $\boldsymbol{\Theta}_\varphi = \mathcal{N}_\epsilon(\gamma_0) \times \boldsymbol{\Theta}_\lambda$, we have

$$\begin{aligned}
E \sup_{\theta \in \boldsymbol{\Theta}_\epsilon} \left| \boldsymbol{\xi}_i'(\boldsymbol{\varphi}) \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1} \boldsymbol{\xi}_i(\boldsymbol{\varphi}) \right| &\leq E \sup_{\theta \in \boldsymbol{\Theta}_\epsilon} \left| Tr[\boldsymbol{\xi}_i(\boldsymbol{\varphi}) \boldsymbol{\xi}_i'(\boldsymbol{\varphi}) \boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}] \right| \\
&\leq E \sup_{\theta \in \boldsymbol{\Theta}_\epsilon} \left\{ \lambda_{\max}[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}] \left\| \boldsymbol{\xi}_i(\boldsymbol{\varphi}) \right\|^2 \right\} \\
&\leq E \sup_{\psi \in \boldsymbol{\Theta}_\psi} \lambda_{\max}[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})^{-1}] E \sup_{\varphi \in \boldsymbol{\Theta}_\varphi} \left\| \boldsymbol{\xi}_i(\boldsymbol{\varphi}) \right\|^2 \\
&\leq E \left(\inf_{\psi \in \boldsymbol{\Theta}_\psi} \lambda_{\min}[\boldsymbol{\Sigma}_\xi(\boldsymbol{\psi})] \right)^{-1} E \sup_{\varphi \in \boldsymbol{\Theta}_\varphi} \left\| \boldsymbol{\xi}_i(\boldsymbol{\varphi}) \right\|^2 \\
&\leq \frac{1}{c_{\min}} E \sup_{\varphi \in \boldsymbol{\Theta}_\varphi} \left\| \boldsymbol{\xi}_i(\boldsymbol{\varphi}) \right\|^2.
\end{aligned}$$

Further

$$\begin{aligned}
E \sup_{\varphi \in \boldsymbol{\Theta}_\varphi} \left\| \boldsymbol{\xi}_i(\boldsymbol{\varphi}) \right\|^2 &= E \sup_{\varphi \in \boldsymbol{\Theta}_\varphi} \left\| \Delta \mathbf{y}_i - \Delta \mathbf{W}_i \boldsymbol{\varphi} \right\|^2 \\
&\leq E \left\| \Delta \mathbf{y}_i \right\|^2 + E \left\| \Delta \mathbf{W}_i \right\|^2 \sup_{\varphi \in \boldsymbol{\Theta}_\varphi} \left\| \boldsymbol{\varphi} \right\|^2.
\end{aligned}$$

But given that $\boldsymbol{\Theta}_\epsilon$ is a compact set $\sup_{\varphi \in \boldsymbol{\Theta}_\varphi} \left\| \boldsymbol{\varphi} \right\|^2$ is bounded. Furthermore, from result (A.2) of Lemma (1) and Liapunov's inequality we have that $E \left\| \Delta \mathbf{y}_i \right\|^2 < K < \infty$

and $E\left\|\Delta\mathbf{W}_i\right\|^2 < K < \infty$. Since $c_{\min}^{-1}$ is bounded by Assumption 7(ii) it follows that $E\sup_{\theta\in\boldsymbol{\Theta}_\epsilon}\left|\boldsymbol{\xi}_i'(\boldsymbol{\varphi})\boldsymbol{\Sigma}_\xi\left(\boldsymbol{\psi}\right)^{-1}\boldsymbol{\xi}_i(\boldsymbol{\varphi})\right| < \infty$ and hence the dominance condition holds.

To establish asymptotic normality of $\hat{\boldsymbol{\theta}}$, by application of the mean value theorem to $\bar{\ell}_N\left(\boldsymbol{\theta}\right)$ around $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, we first note that

$$\bar{\ell}_N\left(\boldsymbol{\theta}\right)-\bar{\ell}_N\left(\boldsymbol{\theta}_0\right)=\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)'\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0)-\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)'\mathbf{H}_N(\bar{\boldsymbol{\theta}})\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right), \tag{A.43}$$

where $\bar{\mathbf{s}}_N(\boldsymbol{\theta})=\partial\bar{\ell}_N\left(\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}$, $\mathbf{H}_N(\boldsymbol{\theta})=-\partial^2\bar{\ell}_N\left(\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, and $\bar{\boldsymbol{\theta}}$ lies on a line segment joining $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. By result (A.27) of Lemma 4, and combining (A.39) and (A.40) we have

$$\begin{aligned}
&\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0)\overset{a.s.}{\to}\mathbf{0},\\
&2\left[\bar{\ell}_N\left(\boldsymbol{\theta}_0\right)-\bar{\ell}_N\left(\boldsymbol{\theta}\right)\right]\overset{a.s.}{\to}\chi\left(\boldsymbol{\psi},\boldsymbol{\psi}_0\right)+\left(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0\right)'\mathbf{A}\left(\boldsymbol{\psi}\right)\left(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0\right)+2\left(\gamma-\gamma_0\right)\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0).
\end{aligned}$$

Hence, in view of (A.43) we must also have

$$\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)'\mathbf{H}_N(\bar{\boldsymbol{\theta}})\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)\overset{a.s.}{\to}\chi\left(\boldsymbol{\psi},\boldsymbol{\psi}_0\right)+\left(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0\right)'\mathbf{A}\left(\boldsymbol{\psi}\right)\left(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0\right)+2\left(\gamma-\gamma_0\right)\kappa(\boldsymbol{\psi},\boldsymbol{\psi}_0). \tag{A.44}$$

But it has aleardy been established by Proposition 3 that on $\boldsymbol{\Theta}_\epsilon$ the right hand side of (A.44) can be equal to zero if and only if $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, and hence we must also have

$$\mathbf{H}_N(\bar{\boldsymbol{\theta}})\overset{a.s.}{\to}\mathbf{H}(\boldsymbol{\theta}_0), \tag{A.45}$$

where $\mathbf{H}(\boldsymbol{\theta}_0)$ *must* be a positive definite matrix given by $\mathbf{H}(\boldsymbol{\theta}_0)=\lim_{N\to\infty}E_0\left[-\partial^2\bar{\ell}_N\left(\boldsymbol{\theta}_0\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\right]$. Applying the mean value theorem to $\bar{\mathbf{s}}_N(\hat{\boldsymbol{\theta}})$ around $\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}_0$ we have

$$\mathbf{0}=\sqrt{N}\bar{\mathbf{s}}_N(\hat{\boldsymbol{\theta}})=\sqrt{N}\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0)-\mathbf{H}_N(\breve{\boldsymbol{\theta}})\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$$

where $\breve{\boldsymbol{\theta}}$ lies on a line segment joining $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Then,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=\mathbf{H}_N^{-1}(\breve{\boldsymbol{\theta}})\left[\sqrt{N}\bar{\mathbf{s}}_N(\boldsymbol{\theta}_0)\right].$$

Since $\breve{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$ is almost surely locally consistent for $\boldsymbol{\theta}_0$ on the set $\boldsymbol{\Theta}_\epsilon$ so is $\breve{\boldsymbol{\theta}}$, and as in (A.45) above $\mathbf{H}_N(\breve{\boldsymbol{\theta}})\overset{a.s.}{\to}\mathbf{H}(\boldsymbol{\theta}_0)$. In addition, using result (A.28) of Lemma 4, we have $\sqrt{N}\bar{\mathbf{s}}_N\left(\boldsymbol{\theta}_0\right)\to_d N\left[\mathbf{0},\mathbf{J}\left(\boldsymbol{\theta}_0\right)\right]$, where $\mathbf{J}\left(\boldsymbol{\theta}_0\right)$ is given by (A.29). Hence

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\to_d N\left(\mathbf{0},\mathbf{V}_\theta\right).$$

where $\mathbf{V}_\theta$ has the familiar sandwich form

$$\mathbf{V}_\theta=\mathbf{H}^{-1}(\boldsymbol{\theta}_0)\mathbf{J}\left(\boldsymbol{\theta}_0\right)\mathbf{H}^{-1}(\boldsymbol{\theta}_0).$$

■

Proof of Proposition 4. Consider the type I error of the test and note that

$$\alpha_N=\Pr\left(\mathcal{X}_N>c_N^2(k)\,|H_0\right)=\Pr\left(\sum_{i=1}^k w_iz_i^2>c_N^2(k)\right),$$

where $z_i \sim IIDN(0,1)$. Now using Lemma A1 of the theory supplement to Chudik et al. (2018) we have

$$\alpha_N = \Pr\left(\sum_{i=1}^{k} w_i z_i^2 > c_N^2(k)\right) \leq \sum_{i=1}^{k} \Pr\left(w_i z_i^2 > k^{-1} c_N^2(k)\right).$$

Therefore, since $w_i > 0$

$$\alpha_N \leq \sum_{i=1}^{k} \Pr\left(z_i^2 > (kw_i)^{-1} c_N^2(k)\right) \leq k \sup_i \Pr\left(z_i^2 > \theta_i^2 c_N^2(k)\right), \tag{A.46}$$

where $\theta_i^2 = (kw_i)^{-1} > 0$. But since $z_i \sim N(0,1)$, then

$$\begin{aligned} \Pr\left(z_i^2 > \theta_i^2 c_N^2(k)\right) &= 1 - \Pr\left(-\theta_i \left|c_N(k)\right| \leq z_i \leq \theta_i \left|c_N(k)\right|\right) \\ &= 2\Phi\left(-\theta_i \left|c_N(k)\right|\right). \end{aligned}$$

Using this result in (A.46) we have

$$\alpha_N \leq 2k \sup_i \Phi\left(-\theta_i \left|c_N(k)\right|\right) = 2k\Phi\left(-\theta_{\min} \left|c_N(k)\right|\right) = 2k\left[1 - \Phi\left(\theta_{\min} \left|c_N(k)\right|\right)\right],$$

where $\theta_{\min}^2 = k^{-1} \inf_i w_i^{-1} = k^{-1} w_1^{-1} > 0$. Hence

$$\Phi\left(\theta_{\min} \left|c_N(k)\right|\right) \leq 1 - \frac{\alpha_N}{2k},$$

and

$$\alpha_N \leq 2k\left[1 - \Phi\left(\theta_{\min} \left|c_N(k)\right|\right)\right] = 2k\Phi\left(-\theta_{\min} \left|c_N(k)\right|\right).$$

Since $\theta_{\min} \left|c_N(k)\right| > 0$, then by (A.1) in Lemma 1 of Bailey et al. (2017, BPS)

$$\Phi\left(-\theta_{\min} \left|c_N(k)\right|\right) \leq (1/2) \exp\left[-\frac{1}{2}\theta_{\min}^2 c_N^2(k)\right],$$

and it follows that

$$\alpha_N \leq k \exp\left[-\frac{1}{2}\theta_{\min}^2 c_N^2(k)\right] = k \exp\left[-\frac{c_N^2(k)}{2kw_1}\right],$$

which ensures that as $N \to \infty$, $\alpha_N \to 0$ so long as $c_N^2(k) \to \infty$.

Also due to the monotonicity property of $\Phi(.)$, we have (for α_N sufficiently small)

$$\theta_{\min} \left|c_N(k)\right| \leq \Phi^{-1}\left(1 - \frac{\alpha_N}{2k}\right),$$

or

$$c_N^2(k) \leq \theta_{\min}^{-2}\left[\Phi^{-1}\left(1 - \frac{\alpha_N}{2k}\right)\right]^2.$$

But by Lemma 3 of BPS we have

$$\left[\Phi^{-1}\left(1 - \frac{\alpha_N}{2k}\right)\right]^2 \leq 2\ln\left(\frac{k}{\alpha_N}\right),$$

and

$$c_N^2(k) \leq 2\theta_{\min}^{-2} \ln\left(\frac{k}{\alpha_N}\right) = 2w_1 k \ln\left(\frac{k}{\alpha_N}\right). \tag{A.47}$$

Consider now the type II error of the test and note that

$$\begin{aligned}
\beta_N &= \Pr\left(\mathcal{X}_N \leq c_N^2(k) \,|H_1\right) = \Pr\left(\sum_{i=1}^{k} w_i \chi_i^2(1, \mu_{i,N}^2) \leq c_N^2(k)\right) \\
&= \Pr\left(\sum_{i=1}^{k} w_i \left(z_i - \mu_{i,N}\right)^2 \leq c_N^2(k)\right).
\end{aligned}$$

Since $w_1 = \max_i(w_i)$, then

$$\sum_{i=1}^{k} w_i \left(z_i - \mu_{i,N}\right)^2 \leq w_1 \sum_{i=1}^{k} \left(z_i - \mu_{i,N}\right)^2,$$

and hence

$$\begin{aligned}
\beta_N &= \Pr\left(\sum_{i=1}^{k} w_i \left(z_i - \mu_{i,N}\right)^2 \leq c_N^2(k)\right) \leq \Pr\left(w_1 \sum_{i=1}^{k} \left(z_i - \mu_{i,N}\right)^2 \leq c_N^2(k)\right) \\
&= \Pr\left(\sum_{i=1}^{k} \left(z_i - \mu_{i,N}\right)^2 \leq \frac{c_N^2(k)}{w_1}\right) \\
&= \Pr\left(\chi^2(k, \mu_N^2) \leq \frac{c_N^2(k)}{w_1}\right),
\end{aligned}$$

where $\chi^2(k, \mu_N^2)$ is a non-central chi-squared random variable with k degrees of freedom and non-centrality parameter, $\mu_N^2 = \sum_{i=1}^{k} \mu_{i,N}^2$. To obtain the rate at which β_N tends to zero with N, we use the normal approximation proposed by Sankaran (1959) for non-central chi-square distributions given by[9]

$$\beta_N \leq \Pr\left(\chi^2(k, \mu_N^2) \leq \frac{c_N^2(k)}{w_1}\right) \approx \Phi\left(\frac{\left(\frac{c_N^2(k)}{w_1\left(k+\mu_N^2\right)}\right)^{h_N} - \left\{1 + h_N A_N \left[h_N - 1 - 0.5(2 - h_N) A_N B_N\right)\right]\right\}}{h_N \sqrt{2A_N}(1 + 0.5 A_N B_N)}\right),$$

where

$$\begin{aligned}
h_N &= 1 - \frac{2}{3}\frac{\left(k + \mu_N^2\right)\left(k + 3\mu_N^2\right)}{\left(k + 2\mu_N^2\right)^2}, \\
A_N &= \frac{k + 2\mu_N^2}{\left(k + \mu_N^2\right)^2}, \qquad B_N = (h_N - 1)(1 - 3h_N).
\end{aligned}$$

Since, k are w_1 are fixed in N, then $A_N = \ominus(\mu_N^{-2})$, $h_N = 1/2 + O(\mu_N^{-2})$, $B_N = 1/4 + O(\mu_N^{-2})$ and it readily follows that as $N \to \infty$, $\beta_N \to 0$ if $c_N^2(k)/\mu_N^2 \to 0$ as $c_N(k)$ and $\mu_N \to \infty$. ■

[9] Also see Patnaik (1949) and Abdel Aty (1964) for other approximations.

Proof of Theorem 2. Consider the event $\{\hat{m} > m_0\}$ where m_0 is the true number of factors. For this event to be true it must be the case that for some $t \in \{1, 2, ..., T-2\}$, at a certain stage in the sequential estimation the null hypothesis of the true number of factors is rejected. That is,

$$\begin{aligned}\Pr(\hat{m} &> m_0) \leq P(\exists t,\ m_0 \text{ is rejected} \,|H_{T-2,t-1}) \\ &\leq \sum_{t=1}^{T-2} \Pr(\mathcal{LR}_N(T-2, t-1) > c^2_{N,T-2,t-1}(k) \,|H_{T-2,t-1}),\end{aligned}$$

where $c^2_{N,T-2,t-1}(k)$ denotes the critical value of the test. For any given t and from Proposition 4 for the type I error of the test we have that as $N \to \infty$

$$\begin{aligned}\alpha_N &= \Pr(\mathcal{LR}_N(T-2, t-1) > c^2_{N,T-2,t-1}(k) \,|H_{T-2,t-1}) \\ &= \Pr\left(\sum_{i=1}^{k} w_i z_i^2 > c^2_{N,T-2,t-1}(k)\right) \to 0,\end{aligned}$$

so long as $c^2_{N,T-2,t-1}(k) \to \infty$, where $z_i \sim IIDN(0, 1)$ from which it follows that

$$\Pr(\hat{m} > m_0) \leq (T-2) \max_{1 \leq t \leq T-2} P\left(\mathcal{LR}_N(T-2, t-1) > c^2_{N,T-2,t-1}(k) \,|H_{T-2,t-1}\right) \to 0. \tag{A.48}$$

Next consider the event $\{\hat{m} < m_0\}$. We have that

$$\begin{aligned}\Pr(\hat{m} &< m_0) = \Pr\left(\max_{1 \leq t \leq T-2} \mathcal{LR}_N(T-2, t-1) \leq c^2_{N,T-2,t-1}(k) \,|H_{T-2,t-1} \text{ is false}\right) \\ &\leq \sum_{t=1}^{T-2} \Pr\left(\mathcal{LR}_N(T-2, t-1) \leq c^2_{N,T-2,t-1}(k) \,|H_{T-2,t-1} \text{ is false}\right).\end{aligned} \tag{A.49}$$

From Proposition 4 for the type II error of the test we have that as $N \to \infty$

$$\begin{aligned}\beta_N &= \Pr\left(\mathcal{LR}_N(T-2, t-1) \leq c^2_{N,T-2,t-1}(k) \,|H_{T-2,t-1} \text{ is false}\right) \\ &= \Pr\left(\sum_{i=1}^{k} w_i \chi_i^2(1, \mu_{i,N}^2) \leq c^2_{N,T-2,t-1}(k)\right) \to 0.\end{aligned}$$

But from (A.49), it readily follows that since $\beta_N \to 0$ as $N \to \infty$, $\Pr(\hat{m} < m_0) \to 0$ which together with (A.48) establishes the desired result. ■

A.3 Derivation of R^2

Consider the panel data model

$$y_{it} = \gamma y_{i,t-1} + \beta x_{it} + \alpha_i + \mu_\delta \delta_t + \zeta_{it}, \quad \zeta_{it} = +\boldsymbol{\eta}_i' \mathbf{f}_t + u_{it},$$

$$x_{it} = \mu_i + \boldsymbol{\vartheta}_i' \mathbf{f}_t + \breve{x}_{it}, \quad \breve{x}_{it} = \rho_x \breve{x}_{i,t-1} + \sqrt{1 - \rho_x^2} \varepsilon_{it},$$

where $\mathbf{f}_t = (f_{1t}, f_{2t}, ..., f_{mt})'$, $\boldsymbol{\eta}_i' = (\eta_{1i}, \eta_{2i}, ..., \eta_{mi})'$, $\boldsymbol{\vartheta}_i = (\vartheta_{1i}, \vartheta_{2i}, ..., \vartheta_{mi})'$, $|\gamma| < 1$ and $|\rho_x| < 1$. To simplify the derivations and without loss of generality we assume that the time effects, δ_t and $\mathbf{f}_t$ are generated with zero means, unit variances and are mutually orthogonal, and ensure that their sample draws over $t = 1, 2, ..., T$ satisfy the following restrictions

$$\begin{aligned} T^{-1}\textstyle\sum_{t=1}^{T} f_{\ell t} &= 0,\ T^{-1}\textstyle\sum_{t=1}^{T} f_{\ell t}^2 = 1,\ T^{-1}\textstyle\sum_{t=1}^{T} f_{\ell t}f_{\ell' t} = 0, \text{ for } \ell \neq \ell'. && \text{(A.50)} \\ T^{-1}\textstyle\sum_{t=1}^{T} \delta_t &= 0,\ T^{-1}\textstyle\sum_{t=1}^{T} \delta_t^2 = 1,\ T^{-1}\textstyle\sum_{t=1}^{T} \delta_t f_{\ell t} = 0, \text{ for all } \ell\ . && \text{(A.51)} \end{aligned}$$

Due to the dependence of x_{it} and ζ_{it} on the same unobserved factors, the regressors and the errors of the above regression are correlated. Following Pesaran and Smith (1994) we base the measurement of R^2 on the following reduced form regressions

$$y_{it} = \tilde{\alpha}_i + \gamma y_{i,t-1} + \beta\breve{x}_{it} + \breve{\zeta}_{it},\ \breve{\zeta}_{it} = \mu_\delta\delta_t + \mathbf{c}_i'\mathbf{f}_t + u_{it}, \tag{A.52}$$

where

$$\tilde{\alpha}_i = \alpha_i + \beta\mu_i \text{ and } \mathbf{c}_i = \beta\boldsymbol{\vartheta}_i + \boldsymbol{\eta}_i. \tag{A.53}$$

It is clear that in (A.52) the regressors, $\breve{x}_{it}$, and the errors, $\breve{\zeta}_{it}$, are uncorrelated and standard formula for R^2 can be used. But to deal with the heterogeneity across the different equations in the panel we use the following average measure of fit

$$R_y^2 = 1 - \frac{N^{-1}\sum_{i=1}^{N} Var(\breve{\zeta}_{it})}{N^{-1}\sum_{i=1}^{N} Var(y_{it})}.$$

Using the above results, and noting that u_{it} and ε_{it} are uncorrelated with δ_t and $\mathbf{f}_t$, then for each unit i we have

$$\begin{aligned} Var(\breve{\zeta}_{it}) &= \mu_\delta^2 Var\left(\delta_t\right) + \mathbf{c}_i' Var(\mathbf{f}_t)\mathbf{c}_i + \sigma^2, \\ Var(y_{it}) &= \frac{\beta^2 Var(\breve{x}_{it}) + Var(\breve{\zeta}_{it})}{1-\gamma^2}. \end{aligned}$$

Under the scaling conditions in (A.50) and (A.51)

$$R_y^2 = \frac{\beta^2 Var(\breve{x}_{it}) + \gamma^2\left[\mu_\delta^2 + \sum_{\ell=1}^{m}\left(N^{-1}\sum_{i=1}^{N} c_{\ell i}^2\right) + \sigma^2\right]}{\beta^2 Var(\breve{x}_{it}) + \mu_\delta^2 + \sum_{\ell=1}^{m}\left(N^{-1}\sum_{i=1}^{N} c_{\ell i}^2\right) + \sigma^2}.$$

It is easily seen that $R_y^2 \geq \gamma^2$ with the equality holding only if $\beta = 0$, namely when an AR(1) specification is considered.

For the unit root case we consider the following average measure of fit

$$R_{\Delta y}^2 = 1 - \frac{N^{-1}\sum_{i=1}^{N} Var(\breve{\zeta}_{it})}{N^{-1}\sum_{i=1}^{N} Var(\Delta y_{it})},$$

where when $\gamma = 1$ (see (A.52) and (A.53))

$$\Delta y_{it} = \tilde{\alpha}_i + \beta\breve{x}_{it} + \breve{\zeta}_{it},$$

with $\tilde{\alpha}_i$, $\breve{x}_{it}$ and $\breve{\zeta}_{it}$ defined as in the case of $\gamma = 1$ in Section 8.1.2. Then

$$\begin{aligned} R^2_{\Delta y} &= 1 - \frac{N^{-1}\sum_{i=1}^{N} Var(\breve{\zeta}_{it})}{\beta^2 N^{-1}\sum_{i=1}^{N} Var(\breve{x}_{it}) + N^{-1}\sum_{i=1}^{N} Var(\breve{\zeta}_{it})} \\ &= \frac{\beta^2 Var(\breve{x}_{it})}{\beta^2 Var(\breve{x}_{it}) + \mu_\delta^2 + \sum_{\ell=1}^{m}\left(N^{-1}\sum_{i=1}^{N} c_{\ell i}^2\right) + \sigma^2}. \end{aligned}$$

A.4 Monte Carlo results for the QML estimator ($T = 5, \gamma = 0.8$)

Table A1 below presents results for the case of $T = 5$ and $\gamma = 0.8$ for the bias, RMSE, size and power of the AR(1) model, including values of N larger than 500, for both Gaussian and non-Gaussian errors. These results show that while over-rejections are observed for smaller sample sizes, size is restored very close to its nominal value as N increases to $2,000$.

Table A1: Bias($\times 100$), RMSE($\times 100$), size(%) and power(%) of γ for the AR(1) model for ($T = 5, \gamma = 0.8$), using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	Gaussian								non-Gaussian							
	Bias ($\times 100$)				RMSE ($\times 100$)				Bias ($\times 100$)				RMSE ($\times 100$)			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
								$\gamma = 0.8$								
100	0.85	0.27	-9.11	-9.67	12.14	12.94	27.50	27.77	0.35	-0.68	-12.14	-13.51	12.42	14.98	30.09	31.68
300	1.21	1.54	-0.67	-0.62	9.46	9.45	14.92	15.06	1.30	1.21	-2.00	-2.67	9.47	10.27	16.04	16.85
500	1.31	1.11	1.03	1.07	7.94	7.93	11.26	11.07	1.45	1.22	-0.35	-0.39	8.12	8.40	12.13	12.47
1,000	0.82	1.31	1.06	1.28	5.94	6.48	8.22	8.19	1.03	0.70	0.75	0.36	6.54	6.44	8.56	9.01
2,000	0.65	0.36	0.65	0.90	4.38	4.21	5.92	6.30	0.44	0.72	0.54	0.39	4.42	4.88	6.39	6.65
	Size				Power				Size				Power			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
	H_0: $\gamma = 0.8$				H_1: $\gamma = 0.7$				H_0: $\gamma = 0.8$				H_1: $\gamma = 0.7$			
100	6.8	8.4	17.9	16.7	21.7	21.8	25.4	24.4	23.3	21.8	26.0	28.5	24.2	25.1	28.2	30.8
300	13.4	12.2	9.4	9.9	34.0	30.2	21.1	21.4	19.0	19.3	15.8	14.7	32.4	30.2	21.6	20.2
500	12.1	11.6	9.4	9.3	39.1	41.5	22.8	22.9	16.9	18.4	11.7	12.4	36.3	39.6	21.3	22.9
1,000	10.0	10.8	9.2	8.6	56.5	53.9	34.5	33.9	14.2	13.3	10.1	11.3	52.9	54.4	30.5	31.9
2,000	6.2	6.3	6.9	7.6	74.2	77.9	51.0	50.9	8.5	8.9	7.5	7.3	74.4	72.1	46.3	46.6

See the notes to Table 3, and Table S3 in the online supplement.

References

Abdel-Aty, S.H. (1964). Approximate formulae for the percentage points and the probability integral of the noncentral χ^2 distribution. *Biometrika*, 41, 538-540.

Ahn, S.C. (2015). Comment on 'IV estimation of panels with factor residuals' by Robertson, D., and Sarafidis, V. *Journal of Econometrics*, 185, 542–544.

Ahn, S. C., and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81, 1203-1227.

Ahn, S. C., and Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics*, 68, 5-27.

Ahn, S. C., Lee,Y.H., and Schmidt, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics*, 101, 219-255.

Ahn, S.C., Y.H. Lee, and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics*, 174, 1-14.

Aiginger, K., Falk, M. (2005). Explaining differences in economic growth among OECD countries. *Empirica*, 32, 19-43.

Anderson, T.W., and Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, 76, 598-606.

Anderson, T.W., and Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18, 47-82.

Arellano, M., and Bond, S.R. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58, 277-297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68, 29-52.

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

Bai, J. (2013). Likelihood approach to dynamic panel models with interactive effects. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2332992.

Bai, J., and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70, 191–221

Bailey, N., Pesaran, M.H. and Smith, L.V. (2017). A multiple testing approach to the regularisation of large sample correlation matrices. unpublished manuscript.

Barro, R. J. (1991). Economic growth in a cross section of countries. *Quarterly Journal of Economics*, 106, 407–443.

Bauer, P., Potscher, B. M., and Hackl, P. (1988). Model selection by multiple test procedures. *Statistics*, 19, 39-44.

Binder, M., Hsiao, C., and Pesaran, M.H. (2005). Estimation and inference in short panel vector autoregressions with unit roots and cointegration. *Econometric Theory*, 21, 795-837.

Blundell, R., and Bond, S.R. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*, 115-143.

Bonferroni, C. E. (1936). Statistical theory of classes and calculation of probabilities. *Pub R Ist Superiore Sci Econ Commerc Firenze*. 8, 36-62.

Caselli, F., Esquivel, G., and Lefort, F. (1996). Reopening the convergence debate: A new look at cross-country growth empirics. *Journal of Economic Growth*, 1, 363–389.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18, 5-46.

Chamberlain, G. (1984). 'Panel Data', in *Handbook of Econometrics*, Volume 2, eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, 1247-1318.

Chudik, A. and Kapetanios, G., and Pesaran, M.H. (2018). A one-covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica*, forthcoming.

Chudik, A., Pesaran, M.H., and Tosetti, E. (2011). Weak and strong cross section dependence and estimation of large panels. *The Econometrics Journal*, 14, C45-C90.

Cragg, J.G., and Donald, S.G. (1997). Inferring the rank of a matrix. *Journal of Econometrics*, 76, 223-250.

Davidson, J. (1994). *Stochastic Limit Theory*, Oxford University Press.

Davidson, J. (2000). *Econometric Theory*, Blackwell, Oxford, UK.

Elhorst, J.P. (2005). Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. *Geographical Analysis*, 37, 85-106.

Feenstra, R.C., Inklaar, R., and Timmer, M.P. (2015). The next generation of the Penn World Table. *American Economic Review,* 105, 3150-3182, available for download at www.ggdc.net/pwt.

Foutz, R. V and Srivastava, R.C (1977). The performance of the likelihood ratio test when the model is incorrect. *The Annals of Statistics*, 5, 1183-1194.

Guariglia A., Rossi M. (2002) Consumption, habit formation and precautionary saving: evidence from the British Household Panel Survey. *Oxford Economic Papers* 54, 1-19.

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

Hayakawa, K. (2016). Identification problem of GMM estimators for panel data models with interactive fixed effects. *Economics Letters*, 139, 22-26.

Hayakawa, K., and Pesaran, M.H. (2015). Robust standard errors in transformed likelihood estimation of dynamic panel data models with cross-sectional heteroskedasticity. *Journal of Econometrics*, 188, 111-134.

Hayashi, K., Bentler, P.M. and Yuan, K-H (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. S*tructural Equation Modeling: A Multidisciplinary Journa*l, 14:3, 505-526.

Heston, A., Summers, R., and Aten, B. (2012). Penn World Table Version 7.1. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.

Holtz-Eakin, D., Newey, W., and Rosen, H.S. (1988). Estimating vector autoregressions with panel data. *Econometrica*, 56, 1371-95.

Hsiao, C., Pesaran, M. H., and Tahmiscioglu, A.K. (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics*, 109, 107-150.

Hsiao, C., and Tahmiscioglu, A.K. (2008). Estimation of dynamic panel data Models with both individual and time specific effects. *Journal of Statistical Planning and Inference*, 138, 2698-2721.

Islam, N. (1995). Growth empirics: A panel data approach. *Quarterly Journal of Economics*, 110, 1127-1170.

Jacobs, J.P.A.M, Ligthart, J.E., and Vrijburg, H. (2009). Dynamic panel data models featuring endogenous interaction and spatially correlated errors. CentER Discussion Paper Series No. 2009-92.

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business & Economic Statistics,* 28, 397–409.

Kapetanios, G., Pesaran, M.H. and T. Yagamata (2011). Panels with nonstationary multifactor error structures. *Journal of Econometrics*, 160, 326-348.

Lee, K., Pesaran, M.H., and Smith, R. (1997). Growth and convergence in a multi-country empirical stochastic Solow model. *Journal of Applied Econometrics,* 12, 357-392.

Lee, K., Pesaran, M.H., and Smith, R. (1998). Growth empirics: A panel data approach - A comment. *Quarterly Journal of Economics,* 113, 319-323.

Lee, L.-F., and Yu, J. (2010). Some recent developments in spatial panel data models. *Regional Science and Urban Economics*, 40, 255-271.

MaCurdy, T.E. (1982). The use of time series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of Econometrics*, 18, 83-114.

Magnus J.R. and Neudecker,H. (1988). Matrix differential calculus with applications in Statistics and Econometrics. John Wiley, Chichester, UK.

Mankiw, N.G., Romer, D., and Weil, D. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics*, CVII, 407-437.

Moral-Benito, E. (2013). Likelihood-based estimation of dynamic panels with predetermined regressors. *Journal of Business & Economic Statistics*, 31, 451-472.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica 46*, 69-85.

Nauges, C., and Thomas, A. (2003). Consistent estimation of dynamic panel data models with time-varying individual effects. *Annales d'Economie et de Statistique*, 70, 53-74.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77, 1447-1479.

Patnaik, P. B. (1949). The noncentral χ^2 and F-distributions and their applications. *Biometrika*, 36, 202-232.

Pesaran, M.H. (2007). A pair-wise approach to testing for output and growth convergence. *Journal of Econometrics*, 138, 312-355.

Pesaran, M.H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74, 967-1012.

Pesaran, M.H. and R. J. Smith (1994). A generalized R^2 criterion for regression models estimated by the instrumental variables method. *Econometrica*, 62, 705-710.

Pesaran, M.H. and Tosetti, E. (2011). Large panels with common factors and spatial correlation. *Journal of Econometrics*, 161, 182-202.

Phillips P. C. B. and Sul, D. (2007). Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *Journal of Econometrics*, 137, 162-188.

Pötscher, B.M. and Prucha, I. (2001). Basic elements of asymptotic theory. In: A Companion to Theoretical Econometrics (B. Baltagi, ed.), Blackwell Publishers.

Prior, M. (2010). You either got it or you don't? The stability of political interest over the lifecycle. Journal of Politics, 72, 746–766.

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

Sala-i-Martin (1996). The classical approach to convergence analysis. *Economic Journal*, 106, 1019-1036.

Sankaran, M. (1959). On the non-central chi-square distribution. *Biometrika*, 46, 235-237.

Sarafidis, V. and Robertson, D. (2009). On the impact of error cross-sectional dependence in short dynamic panel estimation. *The Econometrics Journal*, 12, 62-81.

Sarafidis V. and Wansbeek, T. (2012). Cross-sectional dependence in panel data analysis. *Econometric Reviews*, 31, 483-531.

Satorra, A. and Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In Alexander von Eye and Clifford C. Clogg (Eds.), Latent Variables Analysis: Applications to Developmental Research. (pp. 399-419). SAGE Publications, Inc. : Thousand Oaks, CA.

Su, L. and Z. Yang (2015). QML estimation of dynamic panel data models with spatial errors. *Journal of Econometrics*, 185, 230–258.

Su, L., Y. Zhang, and J. Wei (2016). A practical test for strict exogeneity in linear panel data models with fixed effects, *Economics Letters,* 147, 27-31.

Yuan, K.-H. and Bentler, P.M. (2007). Structural equation modeling, in: C.R. Rao & S. Sinharay (Eds.), Handbook of Statistics 26: Psychometrics (pp. 297-358). Amsterdam: North-Holland.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica,* 50, 1-25.

White, H. (1984). Asymtotic Theory for Econometricians. Academic Press, New York, USA.

Windmeijer, F (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics* 126, 25-51.

An Online Supplement for
Short T Dynamic Panel Data Models with Individual and Interactive Time Effects

Kazuhiko Hayakawa
Hiroshima University

M. Hashem Pesaran
Department of Economics & USC Dornsife INET,
University of Southern California, USA, and Trinity College, Cambridge

L. Vanessa Smith
University of York

September 2018

S.1 Size results for the individual likelihood ratio tests

Tables S1 and S2 below provide size results for testing the individual hypotheses $H_0 : m = m_0 = \{0, 1, 2, ..., T-3\}$ against $H_1 : m = m_{\max} = T-2$ at the 5% significance level using the likelihood ratio statistic $\mathcal{LR}_N(m_{\max}, m_0)$. This statistic is compared to the critical values of the χ^2 distribution with degrees of freedom $DF = T(T+1)/2-(3+Tm-m(m-1)/2)$, where m is equal to its value under the null. Table S2 contains results for the case of Gaussian errors and Table S2 for non-Gaussian errors based on 1,000 replications. Additional results for alternative values of T, N, γ and $\beta = 1$ are available upon request.

Table S1. Size(%) of the individual likelihood ratio tests under Gaussian errors

$H_{m_{\max},m_0}$	$T=5$				$T=10$		
	DF	$AR(1)$	$ARX(1)$		DF	$AR(1)$	$ARX(1)$
$\gamma = 0.4$, $\beta = 1$, $N = 1{,}000$							
$H_{3,0}$	12	5.5	7.3	$H_{8,0}$	52	5.4	7.0
$H_{3,1}$	7	5.3	5.7	$H_{8,1}$	42	4.1	4.8
$H_{3,2}$	3	4.9	5.4	$H_{8,2}$	33	4.2	6.8
				$H_{8,3}$	25	6.2	5.6
				$H_{8,4}$	18	5.4	7.5
				$H_{8,5}$	12	5.3	5.9
				$H_{8,6}$	7	3.4	7.4
				$H_{8,7}$	3	4.5	7.4

Note: $H_{m_{\max},m_0}$ denotes the hypotheses $H_0 : m = m_0 = \{0, 1, 2, ..., T-3\}$ against $H_1 : m = m_{\max} = T-2$. The likelihood ratio statistic is computed as $\mathcal{LR}_N(m_{\max}, m_0) = 2\left[\ell_N\left(\hat{\boldsymbol{\theta}}_{m_{\max}}; m_{\max}\right) - \ell_N\left(\hat{\boldsymbol{\theta}}_{m_0}; m_0\right)\right]$ using the likelihood expression given by (1). All tests are conducted at the 5% significance level. For the data generating process see the notes to Tables 1 and 2.

Table S2. Size(%) of individual likelihood ratio tests under non-Gaussian errors

$H_{m_0 m_{\max}}$	$T=5$		
	DF	$AR(1)$	$ARX(1)$
$\gamma = 0.4$, $\beta = 1$, $N = 1{,}000$			
$H_{3,0}$	12	17.6	16.3
$H_{3,1}$	7	13.5	11.8
$H_{3,2}$	3	9.2	10.1

See the notes to Table S1.

S.2 Monte Carlo results for the ML estimator

The QML estimator reduces to the ML estimator when the errors are generated from a normal distribution as described in the Monte Carlo designs of Sections 8.1.1 and 8.1.2. Here results are presented on the performance of the ML estimator when the number of factors is estimated using the sequential multiple testing likelihood ratio (MTLR) procedure, and when it is known. For these experiments the fixed effects are allowed to be correlated with the errors, and for the ARX(1) model with the regressors as well. In the Monte Carlo designs of HPS this corresponds to $b_1 = b_2 = 1$ with the additional b_0 parameter set to 1 for the ARX(1) model.[10]

[10] For the starting values in the optimization routine used to compute the (Q)ML estimators, we use $\boldsymbol{\theta}_{ini} = (\gamma_{ini}, \omega_{ini}, \sigma^2_{ini}, \boldsymbol{\phi}'_{ini})'$ where $\boldsymbol{\phi}_{ini} = \left(d_{1,ini}, \boldsymbol{\pi}'_{ini}, \boldsymbol{\beta}'_{ini}, \mathbf{d}'_{ini}\right)'$ with $\gamma_{ini} \sim U[-0.999, 0.999]$, $\omega_{ini} \sim U[1, 2]$, $\sigma^2_{ini} \sim U[0.1, 2.1]$ and $\phi_{j,ini} \sim U[-1, 1]$ where $\phi_{j,ini}$ is the jth element of $\boldsymbol{\phi}_{ini}$. In addition ω needs to satisfy $\omega > (T-1)/T$ since $|\boldsymbol{\Omega}| = 1 + T(\omega - 1) > 0$. Specifically, we use five such sets of random starting values and choose the largest among the maximum of the log-likelihood values as the estimate of the (Q)ML estimator.

S.2.1 AR(1)

Simulation results for the AR(1) model are provided in Tables S3-S4. These tables report the bias and RMSE, both multiplied by 100, as well as empirical size and power for the estimator of γ, when the number of factors is estimated and when it is known. The results show that the estimator of γ performs best in terms of RMSE for the true number of factors. However, the differences observed between the true number of factors and the estimated number $\hat{m}$ become minor as N increases. With regard to accuracy of inference, similar to the non-Gaussian case, empirical sizes are close to the nominal level of 0.05 except for the cases where $T = \{5, 10\}$ and $\gamma = 0.8$, for which over-rejections are observed, whether or not the true number of factors is used. As to be expected size distortions decline with T and N. For example, when $T = 10$ we observe size distortions only for $N = 100$ and not when larger values of N are considered. For $T = 5$ we need N to be larger than 500 for size distortions to disappear. See Table A1 in the Appendix which includes additional results for the bias, RMSE, size and power of the AR(1) model for the larger values of N, namely $1,000$ and $2,000$, for both Gaussian and non-Gaussian errors.

Table S3. Bias($\times$100) and RMSE($\times$100) of γ for the AR(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	$T=5$								$T=10$							
	Bias ($\times$100)				RMSE ($\times$100)				Bias ($\times$100)				RMSE ($\times$100)			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
$\gamma = 0.4$																
100	0.31	0.43	-0.04	-0.40	8.60	9.49	12.37	13.19	-0.12	0.01	-0.10	-0.26	4.04	4.10	4.37	4.44
300	-0.18	-0.01	-0.02	-0.08	4.77	4.63	6.43	6.64	-0.01	-0.02	-0.09	-0.02	2.30	2.26	2.50	2.49
500	-0.02	-0.02	-0.01	0.02	3.66	3.71	4.91	5.47	-0.03	-0.07	-0.05	-0.08	1.78	1.84	1.87	1.87
$\gamma = 0.8$																
100	0.85	0.27	-9.11	-9.67	12.14	12.94	27.50	27.77	0.42	0.57	0.54	0.60	5.99	6.06	7.00	6.90
300	1.21	1.54	-0.67	-0.62	9.46	9.45	14.92	15.06	0.16	0.10	0.08	0.24	3.07	3.17	3.68	3.77
500	1.31	1.11	1.03	1.07	7.94	7.93	11.26	11.07	0.04	0.01	0.05	-0.02	2.28	2.32	2.59	2.52

Note: y_{it} is generated as $y_{it} = \alpha_i + \mu_\delta \delta_t + \gamma y_{i,t-1} + \zeta_{it}$, $\zeta_{it} = \boldsymbol{\eta}_i' \mathbf{f}_t + u_{it}$ with the idiosyncratic errors generated as $u_{it} \sim IID\mathcal{N}(0, \sigma^2)$ and $\sigma^2 = 1$, for $i = 1, 2, ..., N; t = -49, 48, ...0, 1, ..., T$, with $y_{i,-50} = \frac{\alpha_i}{1-\gamma} + \sum_{j=0}^{49} \gamma^j \zeta_{i,-j}$. The fixed effects, α_i, are generated as $\alpha_i = \bar{u}_i + v_i$, where $\bar{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$ and $v_i \sim IID\mathcal{N}(0,1)$. The remaining parameters are generated as described in Section 8.1.1. Each f_t is generated once and the same $f_t's$ are used throughout the replications. The first 50 observations are discarded. $\hat{m}$ is the estimated number of factors computed using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50 \frac{p}{(T-2)N}$ and $p = 0.05$. All experiments are based on $2,000$ replications.

Table S4. Size(%) and power(%) of γ for the AR(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	$T=5$								$T=10$							
	Size				Power				Size				Power			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
	H_0: $\gamma = 0.4$				H_1: $\gamma = 0.3$				H_0: $\gamma = 0.4$				H_1: $\gamma = 0.3$			
100	4.8	4.1	5.0	5.6	25.3	25.7	16.6	17.4	5.1	5.4	5.7	5.7	69.5	69.2	64.6	67.5
300	5.6	4.5	5.0	5.1	58.4	57.8	35.4	35.9	4.4	5.0	4.9	4.9	99.4	99.3	98.5	98.5
500	5.5	4.3	4.6	5.1	78.7	78.3	52.6	51.1	4.4	5.9	4.7	4.5	100.0	100.0	100.0	100.0
	H_0: $\gamma = 0.8$				H_1: $\gamma = 0.7$				H_0: $\gamma = 0.8$				H_1: $\gamma = 0.7$			
100	6.8	8.4	17.9	16.7	21.7	21.8	25.4	24.4	8.8	9.2	10.2	9.5	54.6	54.3	49.0	47.3
300	13.4	12.2	9.4	9.9	34.0	30.2	21.1	21.4	4.0	4.7	5.2	5.1	86.9	87.9	82.0	81.4
500	12.1	11.6	9.4	9.3	39.1	41.5	22.8	22.9	4.1	4.4	4.0	4.2	97.1	96.9	95.5	95.5

See the notes to Table S3.

S.2.2 ARX(1)

Simulation results for the ARX(1) model are provided in Tables S5-S6. Similar results as in the AR(1) model are found for the ARX(1) model. Comparing the bias and RMSE values of the γ and β estimators for the case of the true and estimated number of factors, these appear to be very similar and are also very small. With regard to size and power, unlike the AR(1) model, the empirical sizes are close to the nominal level in all cases and power is reasonably high even when the number of factors is estimated.

Table S5. Bias($\times$100) and RMSE($\times$100) of γ and β for the ARX(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	$T=5$								$T=10$							
	Bias ($\times$100)				RMSE ($\times$100)				Bias ($\times$100)				RMSE ($\times$100)			
(m, m_0)	(1,1)	$(\hat{m},1)$	(2,2)	$(\hat{m},2)$	(1,1)	$(\hat{m},1)$	(2,2)	$(\hat{m},2)$	(1,1)	$(\hat{m},1)$	(2,2)	$(\hat{m},2)$	(1,1)	$(\hat{m},1)$	(2,2)	$(\hat{m},2)$
γ, $\gamma=0.4$																
100	0.03	-0.04	-0.02	-0.04	1.46	1.46	1.60	1.57	-0.02	-0.02	0.00	-0.03	0.81	0.81	0.84	0.85
300	0.02	-0.01	-0.03	0.03	0.82	0.83	0.90	0.90	0.02	-0.01	0.00	-0.01	0.45	0.46	0.47	0.48
500	0.02	-0.01	0.00	-0.03	0.63	0.64	0.67	0.68	0.00	0.01	-0.01	0.00	0.35	0.35	0.38	0.38
β																
100	-0.03	-0.02	-0.02	0.08	1.90	1.84	2.03	2.06	-0.02	0.06	-0.02	0.01	1.14	1.14	1.21	1.20
300	-0.02	0.00	0.01	0.02	1.07	1.11	1.18	1.14	-0.01	-0.01	0.01	0.01	0.67	0.66	0.67	0.71
500	0.00	-0.02	-0.02	-0.01	0.83	0.82	0.87	0.90	-0.02	0.00	0.02	0.00	0.50	0.50	0.51	0.52
γ, $\gamma=0.8$																
100	0.05	-0.04	0.02	-0.06	1.90	1.91	1.88	1.88	-0.02	0.00	0.00	-0.02	0.82	0.84	0.82	0.82
300	0.01	0.01	-0.05	0.03	1.07	1.08	1.07	1.06	0.02	0.00	-0.01	0.00	0.46	0.49	0.46	0.47
500	0.03	0.01	0.02	-0.04	0.80	0.83	0.81	0.80	0.00	0.01	-0.01	0.00	0.36	0.36	0.37	0.38
β																
100	-0.04	0.08	-0.03	0.09	3.48	3.27	3.55	3.56	-0.04	-0.04	-0.02	0.11	2.02	2.02	2.09	2.01
300	-0.03	0.03	0.01	0.07	1.98	2.02	2.05	2.01	0.00	0.02	0.02	0.01	1.18	1.16	1.16	1.16
500	0.02	-0.03	-0.03	-0.03	1.52	1.52	1.54	1.57	-0.03	0.00	0.04	-0.01	0.88	0.86	0.88	0.90

Note: y_{it} is generated as $y_{it} = \alpha_i + \mu_\delta \delta_t + \gamma y_{i,t-1} + \beta x_{it} + \zeta_{it}$, $\zeta_{it} = \boldsymbol{\eta}_i' \mathbf{f}_t + u_{it}$ with the idiosyncratic errors generated as $u_{it} \sim IIDN(0, \sigma^2)$ and $\sigma^2 = (1 - R_y^2)/8\left(R_y^2 - \gamma^2\right)$ with $R_y^2 = 0.8$, for $i = 1, 2, ..., N; t = -49, 48, ...0, 1, ..., T$, with $y_{i,-50} = \frac{\alpha_i}{1-\gamma} + \beta \sum_{j=0}^{49} \gamma^j x_{i,-j} + \sum_{j=0}^{49} \gamma^j \zeta_{i,-j}$, and $\beta = 1$. The fixed effects, α_i, are generated as $\alpha_i = \bar{x}_i + \bar{u}_i + v_i$, where $\bar{x}_i = T^{-1} \sum_{t=1}^{T} x_{it}$, $\bar{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$ and $v_i \sim IID\mathcal{N}(0, 1)$. The remaining parameters are generated as described in Section 8.1.2. Each f_t is generated once and the same $f_t's$ are used throughout the replications. The first 50 observations are discarded. $\hat{m}$ is the estimated number of factors computed using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50 \frac{p}{(T-2)N}$ and $p = 0.05$. All experiments are based on 2,000 replications.

Table S6. Size(%) and power(%) of γ and β for the ARX(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0

N	$T=5$								$T=10$							
	Size				Power				Size				Power			
(m, m_0)	(1,1)	$(\hat{m},1)$	(2,2)	$(\hat{m},2)$	(1,1)	$(\hat{m},1)$	(2,2)	$(\hat{m},2)$	(1,1)	$(\hat{m},1)$	(2,2)	$(\hat{m},2)$	(1,1)	$(\hat{m},1)$	(2,2)	$(\hat{m},2)$
γ	H_0: $\gamma=0.40$				H_1: $\gamma=0.38$				H_0: $\gamma=0.40$				H_1: $\gamma=0.38$			
100	5.3	5.2	6.0	5.8	29.9	30.8	27.9	27.9	5.2	5.6	5.4	5.6	71.1	68.0	67.1	68.0
300	4.8	4.2	6.1	6.2	67.1	68.5	62.8	61.6	4.7	5.0	4.8	5.0	99.3	98.8	98.8	98.8
500	4.9	4.7	5.3	4.6	87.5	88.1	84.8	84.8	5.1	5.1	6.0	5.1	100.0	100.0	100.0	100.0
β	H_0: $\beta=1$				H_1: $\beta=0.98$				H_0: $\beta=1$				H_1: $\beta=0.98$			
100	5.0	4.8	5.8	6.0	19.6	18.4	17.3	15.9	5.6	5.4	5.7	5.4	44.2	39.7	40.7	39.7
300	5.1	6.0	5.8	5.5	45.8	46.2	42.4	39.8	5.6	5.6	4.8	5.6	86.6	81.6	83.4	81.6
500	4.8	5.2	4.7	5.1	66.0	67.9	62.5	61.3	4.9	5.1	4.4	4.9	97.9	97.8	96.2	96.9
γ	H_0: $\gamma=0.80$				H_1: $\gamma=0.78$				H_0: $\gamma=0.80$				H_1: $\gamma=0.38$			
100	6.0	5.4	5.8	5.5	19.0	20.7	20.4	22.5	4.1	5.7	5.2	5.7	70.0	70.2	69.3	70.2
300	5.1	4.7	5.7	5.4	46.2	46.4	49.2	47.6	4.5	5.1	4.5	5.1	98.9	98.5	99.3	98.5
500	3.8	5.3	4.5	3.9	65.9	66.4	67.4	71.6	4.5	4.6	5.9	5.1	100.0	99.9	100.0	99.9
β	H_0: $\beta=1$				H_1: $\beta=0.98$				H_0: $\beta=1$				H_1: $\beta=0.98$			
100	5.2	3.8	5.6	5.5	9.4	8.3	9.3	8.9	5.6	4.9	5.8	4.9	18.2	15.2	17.4	15.2
300	5.4	5.5	5.7	5.5	17.0	18.0	16.1	15.5	5.7	5.2	4.5	5.2	42.3	40.5	39.6	40.5
500	4.9	4.5	5.0	4.4	25.7	26.3	24.5	26.7	5.3	4.3	3.9	4.8	63.8	61.8	57.9	59.9

See the notes to Table S5.

S.3 The GMM approach

Let us consider a GMM approach to estimate the dynamic panel data model with interactive effects:

$$y_{it} = \alpha_i + \mathbf{w}_{it}'\boldsymbol{\delta} + \boldsymbol{\lambda}_i'\mathbf{f}_t + \varepsilon_{it}, \qquad (i = 1, 2, ..., N; t = 1, 2, ..., T) \tag{S.1}$$

where $\mathbf{w}_{it} = (y_{i,t-1}, \mathbf{x}_{it}')'$, $\boldsymbol{\delta} = (\gamma, \boldsymbol{\beta}')'$, $\boldsymbol{\lambda}_i = (\lambda_{1i}, ..., \lambda_{mi})'$ and $\mathbf{f}_t = (f_{1t}, ..., f_{mt})'$ are $(m \times 1)$ vectors and ε_{it} are cross-sectionally and temporally uncorrelated. The individual specific effects $\boldsymbol{\lambda}_i$ are allowed to be correlated with $\mathbf{x}_{it}$, while $\mathbf{x}_{it}$ is assumed to be strictly or weakly exogenous. A similar model is considered in Ahn et al. (2013), but there are two differences. The first is that the model under consideration is a dynamic model whereas Ahn et al. (2013) considers a static model. This difference does not cause a serious problem in implementing GMM estimation: minor corrections when selecting the instruments suffice. The second difference is that the current model contains time-invariant fixed effects α_i whereas the model considered in Ahn et al. (2013) does not. Thus the method by Ahn et al. (2013) cannot be applied directly in this case. Hence, we consider two approaches to use the method proposed by Ahn et al. (2013). The first approach is to regard the time-invariant fixed effects as an additional factor to be estimated. The second approach is to take the first-differences prior to applying the quasi-difference approach by Ahn et al. (2013), which is similar to Nauges and Thomas (2003). In the following, we describe each approach.

S.3.1 Approach 1: Quasi-differencing

By incorporating α_i into $\boldsymbol{\lambda}_i'\mathbf{f}_t$ in (S.1), we have the following alternative expression

$$y_{it} = \mathbf{w}_{it}'\boldsymbol{\delta} + \widetilde{\boldsymbol{\lambda}}_i'\widetilde{\mathbf{f}}_t + \varepsilon_{it},$$

where $\widetilde{\boldsymbol{\lambda}}_i = (\alpha_i, \lambda_{1i}, ..., \lambda_{mi})'$ and $\widetilde{\mathbf{f}}_t = (1, f_{1t}, ..., f_{mt})'$. The model in matrix notation can be written as

$$\mathbf{y}_i = \mathbf{W}_i\boldsymbol{\delta} + \widetilde{\mathbf{F}}\widetilde{\boldsymbol{\lambda}}_i + \boldsymbol{\varepsilon}_i, \tag{S.2}$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{iT})'$, $\mathbf{W}_i = (\mathbf{w}_{i1}, ..., \mathbf{w}_{iT})'$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, ..., \varepsilon_{iT})'$ and $\widetilde{\mathbf{F}} = (\widetilde{\mathbf{f}}_1, ..., \widetilde{\mathbf{f}}_T)'$ is a $T \times \tilde{m}$ matrix. Define $\widetilde{\boldsymbol{\Psi}} = \widetilde{\mathbf{F}}\bar{\mathbf{F}}^{-1}$ where $\bar{\mathbf{F}} = (\widetilde{\mathbf{f}}_{T-\tilde{m}+1}, ..., \widetilde{\mathbf{f}}_T)'$. To separately identify $\widetilde{\mathbf{F}}$ from $\widetilde{\boldsymbol{\lambda}}_i$, $\tilde{m}^2$ restrictions are imposed on the factors such that $\widetilde{\mathbf{F}} = (\boldsymbol{\Psi}', \mathbf{I}_{\tilde{m}})'$ where $\boldsymbol{\Psi}$ is a $(T - \tilde{m}) \times \tilde{m}$ matrix of unrestricted parameters obtained as the first $T - \tilde{m}$ rows of $\widetilde{\boldsymbol{\Psi}}$**.** Let $\mathbf{H}_Q$= $(\mathbf{I}_{T-\tilde{m}}, -\boldsymbol{\Psi})'$, so that $\mathbf{H}_Q'\widetilde{\mathbf{F}} = (\mathbf{I}_{T-\tilde{m}}, -\boldsymbol{\Psi})(\boldsymbol{\Psi}', \mathbf{I}_{\tilde{m}})' = \mathbf{0}_{(T-\tilde{m})\times\tilde{m}}$. Then, pre-multiplying equation (S.2) by $\mathbf{H}_Q'$ removes the unobservable effects so that

$$\mathbf{H}_Q'\mathbf{y}_i = \mathbf{H}_Q'\mathbf{W}_i\boldsymbol{\delta} + \mathbf{H}_Q'\boldsymbol{\varepsilon}_i,$$

or

$$\begin{aligned}\dot{\mathbf{y}}_i &= \dot{\mathbf{W}}_i\boldsymbol{\delta} + \boldsymbol{\Psi}\ddot{\mathbf{y}}_i - \boldsymbol{\Psi}\ddot{\mathbf{W}}_i\boldsymbol{\delta} + \dot{\boldsymbol{\varepsilon}}_i - \boldsymbol{\Psi}\ddot{\boldsymbol{\varepsilon}}_i \\ &= \dot{\mathbf{W}}_i\boldsymbol{\delta} + (\mathbf{I}_{T-\tilde{m}} \otimes \ddot{\mathbf{y}}_i')\, vec(\boldsymbol{\Psi}) - \left(vec(\ddot{\mathbf{W}}_i)' \otimes \mathbf{I}_{T-\tilde{m}}\right) vec(\boldsymbol{\delta}' \otimes \boldsymbol{\Psi}) + \dot{\boldsymbol{\varepsilon}}_i - \boldsymbol{\Psi}\ddot{\boldsymbol{\varepsilon}}_i,\end{aligned} \tag{S.3}$$

where $\dot{\mathbf{y}}_i = (y_{i1}, ..., y_{i,T-\tilde{m}})'$, $\ddot{\mathbf{y}}_i = (y_{i,T-\tilde{m}+1}, ..., y_{iT})'$, $\dot{\mathbf{W}}_i = (\mathbf{w}_{i1}, ..., \mathbf{w}_{i,T-\tilde{m}})'$, $\ddot{\mathbf{W}}_i = (\mathbf{w}_{i,T-\tilde{m}+1}, ..., \mathbf{w}_{iT})'$, $\boldsymbol{\Psi}'$= $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{T-\tilde{m}})$, $\dot{\boldsymbol{\varepsilon}}_i = (\varepsilon_{i1}, ..., \varepsilon_{i,T-\tilde{m}})'$, and $\ddot{\boldsymbol{\varepsilon}}_i = (\varepsilon_{i,T-\tilde{m}+1}, ..., \varepsilon_{iT})'$.

The t^{th} equation is given by

$$y_{it} = \boldsymbol{\delta}'\mathbf{w}_{it} + \boldsymbol{\psi}_t'\ddot{\mathbf{y}}_i - \boldsymbol{\psi}_t'\ddot{\mathbf{W}}_i\boldsymbol{\delta} + v_{it}, \qquad (i = 1, ..., N; t = 1, ..., T - \tilde{m}), \tag{S.4}$$

where $v_{it} = (\varepsilon_{it} - \boldsymbol{\theta}_t'\ddot{\boldsymbol{\varepsilon}}_i)$. Since $\mathbf{x}_{it}$ is strictly exogenous, a large number of moment conditions are available. However, since using many instruments causes a large finite sample bias, we consider $(k+1)(T-\tilde{m})(T-\tilde{m}+1)/2+k(T-\tilde{m})\tilde{m}$ moment conditions given by $E[\mathbf{z}_{it}v_{it}] = \mathbf{0}$, for $t = 1, ..., T - \tilde{m}$, where $\mathbf{z}_{it} = (y_{i0}, ..., y_{i,t-1}, \mathbf{x}_{i1}', ..., \mathbf{x}_{it}', \mathbf{x}_{i,T-\tilde{m}+1}', ..., \mathbf{x}_{iT}')'$. In addition to the commonly used instruments $(y_{i0}, ..., y_{i,t-1}, \mathbf{x}_{i1}', ..., \mathbf{x}_{it}')$, we also use $\mathbf{x}_{i,T-\tilde{m}+1}', ..., \mathbf{x}_{iT}'$ as instruments since they are included in the regressor $\ddot{\mathbf{W}}$. In matrix notation the moment conditions can be written as $E\left[\mathbf{Z}_i'\mathbf{v}_i(\boldsymbol{\theta})\right] = \mathbf{0}$, where $\mathbf{Z}_i = diag(\mathbf{z}_{i1}', ..., \mathbf{z}_{i,T-\tilde{m}}')$, $\mathbf{v}_i(\boldsymbol{\theta}) = (v_{i1}, ..., v_{i,T-\tilde{m}})'$ and $\boldsymbol{\theta} = (\boldsymbol{\delta}', \boldsymbol{\psi}')'$ with $\boldsymbol{\psi} = vec(\boldsymbol{\Psi})$.

Then the one-step and two-step GMM estimators are given respectively by

$$\hat{\boldsymbol{\theta}}_{QD1} = \arg\min_{\boldsymbol{\theta}} \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i(\boldsymbol{\theta})'\mathbf{Z}_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{v}_i(\boldsymbol{\theta})\right), \tag{S.5}$$

and

$$\hat{\boldsymbol{\theta}}_{QD2} = \arg\min_{\boldsymbol{\theta}} \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i(\boldsymbol{\theta})'\mathbf{Z}_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{v}_i(\hat{\boldsymbol{\theta}}_{QD1})\mathbf{v}_i(\hat{\boldsymbol{\theta}}_{QD1})'\mathbf{Z}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{v}_i(\boldsymbol{\theta})\right). \tag{S.6}$$

The asymptotic covariance matrix of the above estimators is given, respectively, by

$$Var(\hat{\boldsymbol{\theta}}_{QD1}) = N^{-1}\left(\hat{\mathbf{G}}_{QD1}'\hat{\mathbf{W}}^{-1}\hat{\mathbf{G}}_{QD1}\right)^{-1}\hat{\mathbf{G}}_{QD1}'\hat{\mathbf{W}}^{-1}\hat{\boldsymbol{\Omega}}_{QD1}\hat{\mathbf{W}}^{-1}\hat{\mathbf{G}}_{QD1}\left(\hat{\mathbf{G}}_{QD1}'\hat{\mathbf{W}}^{-1}\hat{\mathbf{G}}_{QD1}\right)^{-1} \tag{S.7}$$

$$Var(\hat{\boldsymbol{\theta}}_{QD2}) = N^{-1}\left(\hat{\mathbf{G}}_{QD2}'\hat{\boldsymbol{\Omega}}_{QD2}^{-1}\hat{\mathbf{G}}_{QD2}\right)^{-1}, \tag{S.8}$$

where $\hat{\mathbf{G}}_j = \partial\bar{\mathbf{g}}(\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}'$ for $j = QD1, QD2$, with $\mathbf{g}_i(\hat{\boldsymbol{\theta}}_j) = \mathbf{Z}_i'\mathbf{v}_i(\hat{\boldsymbol{\theta}}_j)$ and $\bar{\mathbf{g}}(\hat{\boldsymbol{\theta}}_j) = N^{-1}\sum_{i=1}^{N}\mathbf{g}_i(\hat{\boldsymbol{\theta}}_j)$, $\hat{\mathbf{W}} = N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Z}_i$, and $\hat{\boldsymbol{\Omega}}_j = N^{-1}\sum_{i=1}^{N}\mathbf{g}_i(\hat{\boldsymbol{\theta}}_j)\mathbf{g}_i(\hat{\boldsymbol{\theta}}_j)'$. The derivatives involved in $\hat{\mathbf{G}}_j$ are computed numerically.

S.3.2 Approach 2: Quasi-differencing after first-differencing

Taking the first-differences of model (S.1) to remove α_i we have

$$\Delta y_{it} = \Delta\mathbf{w}_{it}'\boldsymbol{\delta} + \boldsymbol{\lambda}_i'\Delta\mathbf{f}_t + \Delta\varepsilon_{it}, \qquad (i = 1, 2, ..., N; t = 2, 3, ..., T)$$

where $\Delta\mathbf{w}_{it} = (\Delta y_{i,t-1}, \Delta\mathbf{x}_{it}')'$, $\boldsymbol{\delta} = (\gamma, \boldsymbol{\beta}')'$, and $\Delta\mathbf{f}_t = \mathbf{f}_t - \mathbf{f}_{t-1}$. The model in matrix notation can be written as

$$\Delta\mathbf{y}_i = \Delta\mathbf{W}_i\boldsymbol{\delta} + \Delta\mathbf{F}\boldsymbol{\lambda}_i + \Delta\boldsymbol{\varepsilon}_i, \tag{S.9}$$

where $\Delta\mathbf{y}_i = (\Delta y_{i2}, ..., \Delta y_{iT})'$, $\Delta\mathbf{W}_i = (\Delta\mathbf{w}_{i2}, ..., \Delta\mathbf{w}_{iT})'$, $\Delta\boldsymbol{\varepsilon}_i = (\Delta\varepsilon_{i2}, ..., \Delta\varepsilon_{iT})'$ and $\Delta\mathbf{F} = (\Delta\mathbf{f}_2, ..., \Delta\mathbf{f}_T)'$ is a $(T-1)\times m$ matrix. Define $\widetilde{\boldsymbol{\Phi}} = \Delta\mathbf{F}\left(\overline{\Delta\mathbf{F}}\right)^{-1}$ where $\overline{\Delta\mathbf{F}} = (\Delta\mathbf{f}_{T-m+1}, ..., \Delta\mathbf{f}_T)'$. To separately identify $\Delta\mathbf{F}$ from $\boldsymbol{\lambda}_i$, m^2 restrictions are imposed on the factors such that $\Delta\mathbf{F} = (\boldsymbol{\Phi}', \mathbf{I}_m)'$ where $\boldsymbol{\Phi}$ is a $(T-1-m)\times m$ matrix of unrestricted parameters obtained as the first $T-1-m$ rows of $\widetilde{\mathbf{\Phi}}$**.** Let $\mathbf{H}_D = (\mathbf{I}_{T-1-m}, -\mathbf{\Phi})'$, so that $\mathbf{H}_D'\Delta\mathbf{F} = (\mathbf{I}_{T-1-m}, -\mathbf{\Phi})(\mathbf{\Phi}', \mathbf{I}_m)' = \mathbf{0}_{(T-1-m)\times m}$. Then, pre-multiplying equation (S.9) by $\mathbf{H}_D'$ removes the unobservable effects so that

$$\mathbf{H}_D'\Delta\mathbf{y}_i = \mathbf{H}_D'\Delta\mathbf{W}_i\boldsymbol{\delta} + \mathbf{H}_D'\Delta\boldsymbol{\varepsilon}_i,$$

or

$$\begin{aligned}\Delta\dot{\mathbf{y}}_i &= \Delta\dot{\mathbf{W}}_i\boldsymbol{\delta} + \mathbf{\Phi}\Delta\ddot{\mathbf{y}}_i - \mathbf{\Phi}\Delta\ddot{\mathbf{W}}_i\boldsymbol{\delta} + \dot{\boldsymbol{\varepsilon}}_i - \mathbf{\Phi}\Delta\ddot{\boldsymbol{\varepsilon}}_i \\ &= \Delta\dot{\mathbf{W}}_i\boldsymbol{\delta} + (\mathbf{I}_{T-1-m}\otimes\Delta\ddot{\mathbf{y}}_i')\, vec(\mathbf{\Phi}) - \left(vec(\Delta\ddot{\mathbf{W}}_i)'\otimes\mathbf{I}_{T-1-m}\right) vec(\boldsymbol{\delta}'\otimes\mathbf{\Phi}) + \Delta\dot{\boldsymbol{\varepsilon}}_i - \mathbf{\Phi}\Delta\ddot{\boldsymbol{\varepsilon}}_i,\end{aligned}$$

where $\Delta\dot{\mathbf{y}}_i = (\Delta y_{i2}, ..., \Delta y_{i,T-m})'$, $\Delta\ddot{\mathbf{y}}_i = (\Delta y_{i,T-m+1}, ..., \Delta y_{iT})'$, $\Delta\dot{\mathbf{W}}_i = (\Delta\mathbf{w}_{i2}, ..., \Delta\mathbf{w}_{i,T-m})'$, $\Delta\ddot{\mathbf{W}}_i = (\Delta\mathbf{w}_{i,T-m+1}, ..., \Delta\mathbf{w}_{iT})'$, $\mathbf{\Phi}' = (\boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_{T-m})$, $\Delta\dot{\boldsymbol{\varepsilon}}_i = (\Delta\varepsilon_{i2}, ..., \Delta\varepsilon_{i,T-m})'$, and $\Delta\ddot{\boldsymbol{\varepsilon}}_i = (\Delta\varepsilon_{i,T-m+1}, ..., \Delta\varepsilon_{iT})'$.

The t^{th} equation is given by

$$\Delta y_{it} = \boldsymbol{\delta}'\Delta\mathbf{w}_{it} + \boldsymbol{\phi}_t'\Delta\ddot{\mathbf{y}}_i - \boldsymbol{\phi}_t'\Delta\ddot{\mathbf{W}}_i\boldsymbol{\delta} + \Delta v_{it}, \qquad (i = 1, ..., N; t = 2, ..., T-m), \quad \text{(S.10)}$$

where $\Delta v_{it} = (\Delta\varepsilon_{it} - \boldsymbol{\phi}_t'\Delta\ddot{\boldsymbol{\varepsilon}}_i)$. Since $\mathbf{x}_{it}$ is strictly exogenous, a large number of moment conditions are available. However, since using many instruments causes a large finite sample bias, we consider $(k+1)(T-1-m)(T-m)/2 + k(T-1-m)m + k(T-1-m)$ moment conditions given by $E[\mathbf{z}_{it}\Delta v_{it}] = \mathbf{0}$, for $t = 2, ..., T-m$, where $\mathbf{z}_{it} = (y_{i0}, ..., y_{i,t-1}, \mathbf{x}_{i0}', \mathbf{x}_{i1}' ..., \mathbf{x}_{it}', \mathbf{x}_{i,T-m+1}', ..., \mathbf{x}_{iT}')'$. In addition to the commonly used instruments $(y_{i0}, ..., y_{i,t-1}, \mathbf{x}_{i0}', ..., \mathbf{x}_{it}')$, we also use $\mathbf{x}_{i,T-m+1}', ..., \mathbf{x}_{iT}'$ as instruments since they are included in the regressor $\Delta\ddot{\mathbf{W}}$. Also, compared to the quasi-difference approach, we additionally use $\mathbf{x}_{i0}$ as instruments. This is because without $\mathbf{x}_{i0}$, the local identification assumption is not satisfied for the linear GMM estimator which is used as the starting value to obtain nonlinear GMM estimators. In matrix notation the moment conditions can be written as $E[\mathbf{Z}_i'\Delta\mathbf{v}_i(\boldsymbol{\theta})] = \mathbf{0}$, where $\mathbf{Z}_i = diag(\mathbf{z}_{i2}', ..., \mathbf{z}_{i,T-m}')$, $\Delta\mathbf{v}_i(\boldsymbol{\theta}) = (\Delta v_{i2}, ..., \Delta v_{i,T-m})'$ and $\boldsymbol{\theta} = (\boldsymbol{\delta}', \boldsymbol{\phi}')'$ with $\boldsymbol{\phi} = vec(\mathbf{\Phi})$.

Then the one-step and two-step GMM estimators are given respectively by

$$\hat{\boldsymbol{\theta}}_{FD1} = \arg\min_{\boldsymbol{\theta}} \left(\frac{1}{N}\sum_{i=1}^N \Delta\mathbf{v}_i(\boldsymbol{\theta})'\mathbf{Z}_i\right)\left(\frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\Delta\mathbf{v}_i(\boldsymbol{\theta})\right), \quad \text{(S.11)}$$

and

$$\hat{\boldsymbol{\theta}}_{FD2} = \arg\min_{\boldsymbol{\theta}} \left(\frac{1}{N}\sum_{i=1}^N \Delta\mathbf{v}_i(\boldsymbol{\theta})'\mathbf{Z}_i\right)\left(\frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\Delta\mathbf{v}_i(\hat{\boldsymbol{\theta}}_{FD1})\Delta\mathbf{v}_i(\hat{\boldsymbol{\theta}}_{FD1})'\mathbf{Z}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N \mathbf{Z}_i'\Delta\mathbf{v}_i(\boldsymbol{\theta})\right). \quad \text{(S.12)}$$

The asymptotic covariance matrix of the above estimators is given, respectively, by

$$Var(\hat{\boldsymbol{\theta}}_{FD1}) = N^{-1}\left(\hat{\mathbf{G}}_{FD1}'\hat{\mathbf{W}}^{-1}\hat{\mathbf{G}}_{FD1}\right)^{-1}\hat{\mathbf{G}}_{FD1}'\hat{\mathbf{W}}^{-1}\hat{\mathbf{\Omega}}_{FD1}\hat{\mathbf{W}}^{-1}\hat{\mathbf{G}}_{FD1}\left(\hat{\mathbf{G}}_{FD1}'\hat{\mathbf{W}}^{-1}\hat{\mathbf{G}}_{FD1}\right)^{-1} \quad \text{(S.13)}$$

$$Var(\hat{\boldsymbol{\theta}}_{FD2}) = N^{-1}\left(\hat{\mathbf{G}}_{FD2}'\hat{\mathbf{\Omega}}_{FD2}^{-1}\hat{\mathbf{G}}_{FD2}\right)^{-1}, \quad \text{(S.14)}$$

where $\hat{\mathbf{G}}_j = \partial\bar{\mathbf{g}}(\hat{\boldsymbol{\theta}}_j)/\partial\boldsymbol{\theta}'$ for $j = FD1, FD2$, with $\mathbf{g}_i(\hat{\boldsymbol{\theta}}_j) = \mathbf{Z}_i'\Delta\mathbf{v}_i(\hat{\boldsymbol{\theta}}_j)$ and $\bar{\mathbf{g}}(\hat{\boldsymbol{\theta}}_j) = N^{-1}\sum_{i=1}^N \mathbf{g}_i(\hat{\boldsymbol{\theta}}_j)$, $\hat{\mathbf{W}} = N^{-1}\sum_{i=1}^N \mathbf{Z}_i'\mathbf{Z}_i$, and $\hat{\mathbf{\Omega}}_j = N^{-1}\sum_{i=1}^N \mathbf{g}_i(\hat{\boldsymbol{\theta}}_j)\mathbf{g}_i(\hat{\boldsymbol{\theta}}_j)'$. The derivatives involved in $\hat{\mathbf{G}}_j$ are computed numerically.

S.3.3 Starting values

For the computation of the above nonlinear GMM estimators, starting values are required. We use the linear GMM estimator by Hayakawa (2012) as the starting value. This can reduce the computational time compared with the case where several random starting values are used.

To define the linear GMM estimator, let us define $L_1 = L_2 = 1$ for $\tilde{m} = 1$, and $\mathbf{L}_1 = (\mathbf{I}_{\tilde{m}}, \mathbf{0}_{\tilde{m}})$ and $\mathbf{L}_2 = (\mathbf{0}_{\tilde{m}}, \mathbf{I}_{\tilde{m}})$ for $\tilde{m} > 1$. Also, define $\check{\mathbf{y}}_i = (y_{i,T-\tilde{m}}, y_{i,T-\tilde{m}+1}, .., y_{iT})' = (y_{i,T-\tilde{m}}, \ddot{\mathbf{y}}_i')'$. Then, noting that $\ddot{\mathbf{W}}_i = \left(\ddot{\mathbf{y}}_{i,-1}, \ddot{\mathbf{X}}_{it}\right)$ where $\ddot{\mathbf{y}}_{i,-1}= (y_{i,T-\tilde{m}}, y_{i,T-\tilde{m}+1}, .., y_{iT-1})'$, $\ddot{\mathbf{y}}_i = \mathbf{L}_2\check{\mathbf{y}}_i$ and $\ddot{\mathbf{y}}_{i,-1} = \mathbf{L}_1\check{\mathbf{y}}_i$, we have

$$
\begin{aligned}
\dot{\mathbf{y}}_i &= \dot{\mathbf{W}}_i\boldsymbol{\delta} + \boldsymbol{\Psi}\ddot{\mathbf{y}}_i - \boldsymbol{\Psi}\ddot{\mathbf{W}}_i\boldsymbol{\delta} + \dot{\boldsymbol{\varepsilon}}_i - \boldsymbol{\Psi}\ddot{\boldsymbol{\varepsilon}}_i \\
&= \dot{\mathbf{W}}_i\boldsymbol{\delta} + \boldsymbol{\Psi}\mathbf{L}_2\check{\mathbf{y}}_i - \boldsymbol{\Psi}\left(\gamma\mathbf{L}_1\check{\mathbf{y}}_i + \ddot{\mathbf{X}}_i\boldsymbol{\beta}\right) + \dot{\boldsymbol{\varepsilon}}_i - \boldsymbol{\Psi}\ddot{\boldsymbol{\varepsilon}}_i \\
&= \dot{\mathbf{W}}_i\boldsymbol{\delta} + \boldsymbol{\Psi}\left(\mathbf{L}_2 - \gamma\mathbf{L}_1\right)\check{\mathbf{y}}_i - \boldsymbol{\Psi}\ddot{\mathbf{X}}_i\boldsymbol{\beta} + \mathbf{v}_i \\
&= \dot{\mathbf{W}}_i\boldsymbol{\delta} + \boldsymbol{\Upsilon}\check{\mathbf{y}}_i - \boldsymbol{\Psi}\ddot{\mathbf{X}}_i\boldsymbol{\beta} + \mathbf{v}_i \\
&= \dot{\mathbf{W}}_i\boldsymbol{\delta} + \left(\mathbf{I}_{T-\tilde{m}} \otimes \check{\mathbf{y}}_i'\right) vec(\boldsymbol{\Upsilon}') - \left(vec(\ddot{\mathbf{X}}_i)' \otimes \mathbf{I}_{T-\tilde{m}}\right) vec(\boldsymbol{\beta}' \otimes \boldsymbol{\Psi}) + \mathbf{v}_i \\
&= \widetilde{\mathbf{X}}_i\boldsymbol{\pi} + \mathbf{v}_i
\end{aligned}
$$

where $\boldsymbol{\Upsilon} = \boldsymbol{\Psi}\left(\mathbf{L}_2 - \gamma\mathbf{L}_1\right)$, $\mathbf{X}_i = \left(\dot{\mathbf{W}}_i, \left(\mathbf{I}_{T-\tilde{m}} \otimes \check{\mathbf{y}}_i'\right), -\left(vec(\ddot{\mathbf{X}}_i)' \otimes \mathbf{I}_{T-\tilde{m}}\right)\right)$ and $\boldsymbol{\pi} = \left(\boldsymbol{\delta}', vec(\boldsymbol{\Upsilon}')', vec(\boldsymbol{\beta}' \otimes \boldsymbol{\Psi})'\right)' = (\boldsymbol{\pi}_1', \boldsymbol{\pi}_2', \boldsymbol{\pi}_3')'$ with $\boldsymbol{\pi}_1 = \boldsymbol{\delta}, \boldsymbol{\pi}_2 = vec(\boldsymbol{\Upsilon}')$, $\boldsymbol{\pi}_3 = vec(\boldsymbol{\beta}' \otimes \boldsymbol{\Psi})$. We consider this particular model rather than the original model (S.3) because perfect multicollinearity between $\ddot{\mathbf{y}}_i$ and $\ddot{\mathbf{W}}_i$ occurs in (S.3) when $\tilde{m} > 1$. Since this is a linear model in $\boldsymbol{\pi}$ with moment conditions $E\left[\mathbf{Z}_i'\mathbf{v}_i(\boldsymbol{\pi})\right] = \mathbf{0}$, a closed form solution is obtained as

$$
\begin{aligned}
\hat{\boldsymbol{\pi}} &= \left[\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\widetilde{\mathbf{X}}_i\right)\right]^{-1} \\
&\quad\times\left[\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i'\dot{\mathbf{y}}_i\right)\right].
\end{aligned}
$$

Hence, $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$ are consistent estimates of $\boldsymbol{\delta}$ and $vec(\boldsymbol{\Upsilon}')$, respectively. To recover $\boldsymbol{\Psi}$ from the estimate of $\boldsymbol{\Upsilon}$, since

$$
vec\left(\boldsymbol{\Upsilon}'\right) = vec\left(\left(\mathbf{L}_2 - \gamma\mathbf{L}_1\right)'\boldsymbol{\Psi}'\right) = \left(\mathbf{I}_{T-\tilde{m}} \otimes \left(\mathbf{L}_2 - \gamma\mathbf{L}_1\right)'\right) vec\left(\boldsymbol{\Psi}'\right) = \mathbf{A}vec\left(\boldsymbol{\Psi}'\right),
$$

$vec\left(\boldsymbol{\Psi}'\right)$ is obtained as $vec\left(\boldsymbol{\Psi}'\right) = \left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'vec\left(\boldsymbol{\Upsilon}'\right)$. In the computation of the nonlinear GMM estimators, estimates of $\boldsymbol{\delta}$ and $vec\left(\boldsymbol{\Psi}'\right)$ are obtained from $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$ and are used as the starting values of the numerical optimization.

The same procedure can be used in approach 2 by replacing the $\mathbf{y}_i$'s and $\mathbf{W}_i$'s with their first differences.

S.3.4 The AR(1) model

Estimation of the AR(1) model is exactly the same as above after removing all $\mathbf{x}$'s from both the model and instruments. However, for the starting value, we cannot use the linear estimator since the number of moment conditions is always smaller than that of the unknown reduced form parameters. Hence in the Monte Carlo simulations for this case we use random starting values. Specifically, we use

$$\gamma_{ini} \sim U\left(-1,1\right), \ \psi_{j,ini} \sim U\left(-2,2\right), \quad (j=1,...,(T-\tilde{m})\tilde{m})$$

for approach 1 and

$$\gamma_{ini} \sim U\left(-1,1\right), \ \psi_{j,ini} \sim U\left(-2,2\right), \quad (j=1,...,(T-1-m)m)$$

for approach 2.

S.4 Monte Carlo results for the ML and GMM estimators

Tables S7-S10 present results on the bias, RMSE, size and power for the ML and the GMM quasi-difference (QD) and first-difference (FD) estimators of ALS, when the number of factors is assumed to be known and the errors in the Monte Carlo designs of HPS are generated as normal. In these experiments the fixed effects are not correlated with the errors, only with the regressors for the ARX(1) model, as this would render the GMM estimators inconsistent. This is equivalent to setting $b_1 = 0$ and $b_2 = 1$, with the additional b_0 parameter set to 1 for the ARX(1) model in the Monte Carlo designs of Sections 8.1.1 and 8.1.2. The GMM estimators are computed as shown in Section S.3. Results for the AR(1) model are presented in Tables S7 and S8 and in Tables S9 and S10 for the ARX(1) model.

Table S7. Bias($\times$100) and RMSE($\times$100) of γ for the QML and GMM estimators in the case of the AR(1) model, using the true number of factors, m_0

N	$T=10$																				
	Bias ($\times$100)					RMSE ($\times$100)					Bias ($\times$100)					RMSE ($\times$100)				
	ML	GMM				ML	GMM				ML	GMM				ML	GMM			
		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2
m_0	1										2									
$\gamma=0.4$																				
100	0.42	11.41	15.04	-34.15	-22.96	3.96	36.97	36.10	35.50	25.05	-0.21	49.58	49.44	-26.63	-24.84	4.21	52.74	53.24	58.14	50.26
300	-0.07	-0.88	4.37	-18.33	-7.55	2.35	24.01	22.50	19.41	9.28	-0.08	48.33	47.33	-12.64	-12.08	2.44	52.21	52.23	43.45	34.81
500	-0.02	-0.87	3.98	-13.00	-4.10	1.79	20.31	19.49	14.13	5.79	-0.02	46.85	45.83	-7.61	-8.42	1.85	51.43	51.44	36.93	28.96
$\gamma=0.8$																				
100	0.69	13.17	13.18	-70.69	-62.14	6.12	20.93	21.64	72.10	64.84	0.30	17.60	17.51	-71.91	-68.08	6.63	18.61	18.64	90.92	87.27
300	0.11	9.27	9.93	-51.59	-34.78	3.19	20.13	19.68	52.68	37.01	0.10	17.48	17.40	-49.83	-42.84	3.53	18.52	18.42	71.04	63.46
500	0.08	8.26	9.46	-42.71	-24.51	2.34	19.05	18.04	43.86	26.77	0.06	16.75	16.74	-41.85	-34.15	2.54	18.44	18.27	63.02	55.19

Note: GMM QD1, QD2, FD1 and FD2 are the quasi-difference and first- difference ALS one step and two step estimators respectively computed as described in Section S.3. All experiments are based on 2,000 replications. See also the notes to Table S3.

Table S8. Size(%) and power(%) of γ for the QML and GMM estimators in the case of the AR(1) model, using the true number of factors, m_0

N	$T=10$																				
	Size					Power					Size					Power					
	ML	GMM				ML	GMM				ML	GMM				ML	GMM				
		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2	
m_0	1										2										
γ	H_0: $\gamma=0.40$					H_1: $\gamma=0.34$					H_0: $\gamma=0.40$					H_1: $\gamma=0.34$					
100	5.3	65.3	71.3	89.4	88.7	31.7	78.1	84.0	96.6	97.1	5.0	86.4	88.1	30.6	42.9	30.1	86.0	88.0	31.6	45.9	
300	5.4	44.9	36.2	73.1	46.2	74.2	71.6	70.1	92.5	85.4	4.9	82.9	83.4	19.7	32.7	70.1	82.7	83.2	19.4	32.3	
500	5.3	38.5	28.2	60.8	27.6	91.6	72.2	70.3	89.6	82.1	3.2	79.4	81.0	14.8	25.5	89.4	79.3	81.5	14.3	28.4	
γ	H_0: $\gamma=0.80$					H_1: $\gamma=0.74$					H_0: $\gamma=0.80$					H_1: $\gamma=0.74$					
100	9.0	93.8	96.4	98.9	99.7	31.1	95.0	97.0	99.1	100.0	8.3	95.6	96.9	51.8	65.0	27.5	94.0	96.2	54.0	73.4	
300	5.1	87.3	88.8	98.4	95.8	53.9	91.5	92.9	99.5	99.0	4.6	95.4	95.8	35.7	48.2	47.1	93.9	95.4	37.5	61.2	
500	4.6	84.4	83.0	96.8	88.5	72.9	89.5	88.3	98.9	97.1	3.7	93.0	93.8	29.4	41.9	64.9	91.3	93.8	30.5	59.7	

See the notes to Table S7.

Table S9. Bias($\times$100) and RMSE($\times$100) of γ and β for the QML and GMM estimators in the case of the ARX(1) model, using the true number of factors, m_0

N	$T=5$										$T=10$									
	Bias ($\times$100)					RMSE ($\times$100)					Bias ($\times$100)					RMSE ($\times$100)				
	ML	GMM				ML	GMM				ML	GMM				ML	GMM			
		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2
m_0	1																			
γ, $\gamma=0.4$																				
100	0.03	0.36	0.13	0.08	0.12	1.46	3.13	3.09	4.05	2.76	-0.02	10.21	10.12	-2.43	-2.09	0.81	20.64	20.55	3.22	2.82
300	0.02	0.48	0.16	-0.01	0.04	0.82	1.37	1.30	2.44	1.60	0.02	2.14	1.88	-0.85	-0.35	0.45	8.46	8.11	1.89	0.83
500	0.02	0.49	0.16	0.05	0.05	0.63	1.11	1.02	1.88	1.21	0.00	0.95	0.71	-0.55	-0.16	0.35	4.28	3.98	1.82	0.53
β																				
100	-0.03	-0.04	0.00	0.20	0.08	1.90	2.40	2.38	4.45	3.21	-0.02	-8.93	-8.85	-0.67	-0.52	1.14	20.01	19.84	4.14	3.55
300	-0.02	-0.13	-0.03	0.18	0.06	1.07	1.37	1.35	2.80	1.87	-0.01	-1.60	-1.46	-0.08	0.05	0.67	8.22	7.60	3.12	1.19
500	0.00	-0.14	-0.04	0.09	0.03	0.83	1.03	1.00	2.06	1.42	-0.02	-0.50	-0.45	-0.09	-0.01	0.50	4.30	3.81	2.56	0.75
γ, $\gamma=0.8$																				
100	0.05	8.62	7.12	-0.47	-0.37	1.90	10.17	9.26	5.11	4.28	-0.02	12.47	12.40	-5.74	-4.98	0.82	12.50	12.43	7.20	6.23
300	0.01	8.50	5.57	-0.03	-0.10	1.07	9.71	7.82	2.98	2.39	0.02	12.29	11.80	-1.95	-0.78	0.46	12.30	11.81	3.04	1.30
500	0.03	8.56	5.03	0.06	0.02	0.80	9.68	7.28	2.30	1.84	0.00	12.23	11.66	-1.27	-0.35	0.36	12.23	11.66	2.07	0.73
β																				
100	-0.04	-1.75	-0.37	-0.27	-0.08	3.48	7.17	5.70	7.47	6.17	-0.04	-14.16	-13.64	-4.46	-3.83	2.02	15.35	14.84	10.31	8.90
300	-0.03	-0.29	0.03	-0.15	-0.07	1.98	3.31	2.90	4.40	3.41	0.00	-12.45	-8.75	-1.06	-0.32	1.18	12.94	9.35	5.90	2.49
500	0.02	-0.06	0.02	-0.22	-0.09	1.52	2.31	2.08	3.28	2.64	-0.03	-11.92	-7.64	-0.71	-0.18	0.88	12.22	8.02	4.47	1.46
m_0	2																			
γ, $\gamma=0.4$																				
100	-0.02	0.51	0.34	-1.91	-1.61	1.60	4.87	3.98	7.42	7.13	0.00	-0.58	-0.46	-1.27	-0.98	0.84	4.34	4.24	5.28	3.85
300	-0.03	0.01	-0.09	-0.63	-0.49	0.90	1.80	1.67	3.78	3.43	0.00	-0.35	-0.16	-0.25	-0.08	0.47	1.83	1.55	4.27	1.45
500	0.00	-0.10	-0.11	-0.30	-0.24	0.67	1.25	1.25	2.49	2.13	-0.01	-0.27	-0.10	0.10	0.01	0.38	1.13	0.89	4.34	0.96
β																				
100	-0.02	-0.12	-0.03	-2.15	-1.83	2.03	3.18	3.26	13.32	13.41	-0.02	0.26	0.17	0.09	0.06	1.21	4.39	4.23	5.24	4.01
300	0.01	0.02	0.12	-0.59	-0.36	1.18	1.77	1.76	6.89	6.43	0.01	0.22	0.13	0.02	0.02	0.67	1.53	1.08	3.81	1.53
500	-0.02	0.04	0.07	-0.11	0.00	0.87	1.35	1.31	3.78	3.41	0.02	0.17	0.08	0.07	0.03	0.51	1.06	0.80	3.23	1.02
γ, $\gamma=0.8$																				
100	0.02	7.87	5.78	-17.82	-17.32	1.88	9.54	8.29	31.56	32.49	0.00	-0.61	-0.48	-3.66	-2.86	0.82	3.91	3.81	7.26	5.48
300	-0.05	4.49	2.58	-7.65	-7.24	1.07	7.13	5.59	18.26	19.05	-0.01	-1.16	-0.79	-1.37	-0.69	0.46	1.59	1.25	4.41	1.79
500	0.02	2.82	1.54	-3.32	-3.07	0.81	5.71	4.16	11.75	12.11	-0.01	-0.95	-0.59	-0.86	-0.38	0.37	1.22	0.89	3.51	1.11
β																				
100	-0.03	0.26	0.56	-19.87	-19.61	3.55	6.93	6.93	37.22	38.28	-0.22	0.63	0.52	-0.77	-0.55	2.09	3.99	3.80	10.32	7.96
300	0.01	0.70	0.37	-9.45	-8.59	2.05	3.92	3.47	22.59	22.89	0.02	0.58	0.36	-0.38	-0.03	1.16	2.09	1.75	6.93	2.90
500	-0.03	0.53	0.17	-4.34	-3.85	1.54	2.90	2.42	14.59	14.59	0.04	0.45	0.26	-0.08	0.04	0.88	1.60	1.34	5.17	1.83

Note: GMM QD1, QD2, FD1 and FD2 are the quasi-difference and first- difference ALS one step and two step estimators respectively computed as described in Section S.3. All experiments are based on 2,000 replications. See also the notes to Table S5.

Table S10. Size(%) and power(%) of γ and β for the QML and GMM estimators in the case of the ARX(1) model, using the true number of factors, m_0

N	$T=5$										$T=10$									
	Size					Power					Size					Power				
	ML	GMM				ML	GMM				ML	GMM				ML	GMM			
		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2		QD1	QD2	FD1	FD2
m_0	1																			
γ	H_0: $\gamma=0.40$					H_1: $\gamma=0.38$					H_0: $\gamma=0.40$					H_1: $\gamma=0.38$				
100	5.3	6.1	14.8	3.6	12.4	29.9	14.1	30.6	8.2	19.7	5.2	31.5	87.8	28.3	82.8	71.2	33.5	87.4	69.7	97.7
300	4.8	6.6	9.6	3.8	7.2	67.1	22.5	39.9	16.5	29.7	4.7	16.8	45.5	12.6	26.2	99.3	27.0	77.7	73.6	97.4
500	4.9	6.9	8.5	3.9	6.3	87.5	31.3	55.4	23.1	41.2	5.1	16.5	37.5	10.1	14.3	100.0	34.3	88.7	84.0	99.6
β	H_0: $\beta=1$					H_1: $\beta=0.98$					H_0: $\beta=1$					H_1: $\beta=0.98$				
100	5.0	6.0	15.1	4.2	11.1	19.6	15.3	28.3	7.3	19.4	5.6	27.5	84.6	7.8	75.1	44.1	33.0	88.5	13.2	81.1
300	5.1	5.9	10.6	4.9	7.9	45.9	36.9	44.8	11.9	24.3	5.6	10.6	35.3	5.0	27.2	86.6	30.4	81.1	15.1	73.6
500	4.9	5.0	8.8	4.3	7.4	66.0	54.9	61.6	17.3	33.0	4.9	9.0	27.4	4.7	15.9	97.9	43.3	91.5	20.2	90.1
γ	H_0: $\gamma=0.80$					H_1: $\gamma=0.78$					H_0: $\gamma=0.80$					H_1: $\gamma=0.78$				
100	6.0	76.5	69.2	4.7	13.0	19.0	77.5	70.8	8.4	19.5	4.1	100.0	100.0	55.4	95.5	70.0	100.0	100.0	77.6	99.2
300	5.1	76.7	57.1	4.0	7.9	46.2	78.3	58.1	10.4	20.2	4.5	100.0	100.0	20.2	41.0	98.9	100.0	100.0	58.3	95.9
500	3.8	77.5	54.7	3.9	7.4	65.9	79.5	54.4	12.8	23.1	4.5	100.0	100.0	13.7	21.4	100.0	100.0	100.0	59.5	98.9
β	H_0: $\beta=1$					H_1: $\beta=0.98$					H_0: $\beta=1$					H_1: $\beta=0.98$				
100	5.2	12.4	18.8	4.4	13.6	9.4	15.7	21.6	5.8	15.9	5.6	85.4	98.5	15.3	80.0	18.2	91.1	99.4	21.3	82.7
300	5.4	7.5	12.4	4.5	8.3	17.0	12.8	18.9	7.8	14.4	5.7	99.0	97.5	7.6	33.3	42.3	99.8	99.7	12.3	53.4
500	4.9	5.3	9.6	4.6	8.2	25.7	14.5	23.6	9.4	17.0	5.3	99.9	98.8	6.3	18.0	63.8	100.0	99.9	12.7	56.0
m_0	2																			
γ	H_0: $\gamma=0.40$					H_1: $\gamma=0.38$					H_0: $\gamma=0.40$					H_1: $\gamma=0.38$				
100	6.0	5.5	12.6	7.7	11.7	27.9	10.8	19.8	12.1	19.1	5.4	12.7	55.6	9.5	57.4	67.1	51.8	89.4	27.8	80.7
300	6.1	4.0	9.3	4.9	8.1	62.8	23.5	33.9	15.0	25.2	4.8	8.2	20.0	6.1	27.4	98.8	77.0	94.2	23.3	82.3
500	5.3	3.8	7.9	4.9	7.2	84.8	38.8	47.3	20.9	37.2	6.0	7.4	15.9	5.4	20.1	100.0	88.3	98.1	22.0	92.7
β	H_0: $\beta=1$					H_1: $\beta=0.98$					H_0: $\beta=1$					H_1: $\beta=0.98$-				
100	5.8	4.8	12.6	5.9	11.0	17.3	10.7	21.6	8.0	15.0	5.7	7.9	49.4	9.2	62.2	40.7	12.8	60.5	12.8	66.7
300	5.8	4.6	8.9	5.0	6.5	42.4	20.1	26.0	10.0	16.9	4.8	5.7	19.3	6.0	28.5	83.4	29.0	68.4	13.7	62.2
500	4.7	4.3	7.2	3.3	5.9	62.5	29.6	38.6	11.9	23.6	4.5	5.3	15.5	5.4	18.9	96.2	44.1	84.8	12.9	76.6
γ	H_0: $\gamma=0.80$					H_1: $\gamma=0.78$					H_0: $\gamma=0.80$					H_1: $\gamma=0.78$				
100	5.8	74.5	56.0	29.7	34.5	20.4	74.1	57.2	31.9	37.2	5.2	27.4	68.5	22.6	76.5	69.2	68.7	94.6	38.9	89.1
300	5.7	44.5	30.5	20.8	20.6	49.2	49.6	40.8	24.4	29.3	4.5	24.4	37.1	9.0	38.9	99.3	90.4	97.1	25.3	86.6
500	4.5	29.9	19.5	12.6	13.5	67.4	42.5	37.0	19.0	30.7	5.9	25.4	32.1	6.0	29.3	100.0	96.3	99.1	22.0	93.4
β	H_0: $\beta=1$					H_1: $\beta=0.98$					H_0: $\beta=1$					H_1: $\beta=0.98$				
100	5.6	7.1	16.4	26.7	31.4	9.3	8.1	18.1	28.1	33.9	5.8	7.4	48.7	10.7	65.7	17.3	8.2	51.7	12.1	67.2
300	5.7	7.7	12.0	19.4	19.7	16.1	10.0	14.8	21.0	24.0	4.4	5.6	17.7	7.6	31.5	39.6	10.6	34.5	10.1	45.0
500	5.0	6.8	8.8	12.9	12.9	24.5	12.2	19.8	15.2	19.1	3.9	5.6	14.0	5.7	21.5	57.9	15.2	45.0	8.3	46.4

See the notes to Table S9.

S.5 Monte Carlo results for the unit root case ($\gamma = 1$)

S.5.1 Performance of the sequential multiple testing likelihood ratio procedure

Table S11 provides results on the performance of the sequential MTLR procedure in the unit root case, for the AR(1) and the ARX(1) models, respectively. Specifically they report the number of times, in percent, that the estimated number of factors, $\hat{m}$, based on the sequential MTLR procedure outlined in Section 7.1 is equal to the true number of factors m_0. The sequential MTLR procedure is implemented using the $\mathcal{LR}_N(m_{\max}, m_0)$ statistic for testing $m = m_0 = \{0, 1, 2, .., T-3\}$ against $m = m_{\max} = T-2$, with significance level $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$, using the critical values of the chi-square distribution with degrees of freedom given by (44). Results are reported for the case of both Gaussian and non-Gaussian errors. The tables show that the sequential MTLR procedure works very well even for the unit root case. The performance is very similar to the stationary case with $|\gamma| < 1$, and indeed, the probability of selecting the true number of factors exceeds 95% in most cases even under non-Gaussianity.

Table S11. Empirical frequency of correctly selecting the true number of factors, m_0, using the sequential MTLR procedure in the unit root case ($\gamma = 1$)

N	$T=5$						$T=10$					
	$m_0=0$		$m_0=1$		$m_0=2$		$m_0=0$		$m_0=1$		$m_0=2$	
	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian	Gaussian	non-Gaussian
AR(1)												
100	99.0	96.2	99.2	96.7	97.4	94.7	99.7	95.7	99.0	95.9	99.5	96.6
300	99.7	97.5	99.5	98.1	97.7	97.2	100.0	98.6	99.8	98.9	99.8	99.6
500	99.8	98.9	100.0	98.9	98.3	97.6	100.0	98.7	99.9	99.4	100.0	99.5
1,000	99.9	98.9	99.9	99.8	99.1	97.9	100.0	99.2	99.9	99.5	100.0	99.7
ARX(1)												
100	98.8	94.5	98.7	95.1	98.5	96.1	99.4	93.5	98.6	94.9	99.1	95.1
300	99.4	97.7	99.5	98.5	99.7	99.1	99.6	97.5	99.7	98.8	99.7	99.0
500	99.7	97.9	99.6	98.1	99.6	98.7	100.0	98.2	100.0	98.7	99.9	99.3
1,000	100.0	99.4	100.0	99.6	99.9	99.6	100.0	99.2	99.9	99.3	100.0	99.3

Note: The first-differences are initially generated and then cumulated to obtain y_{it} starting from any arbitrary value for y_{i0}. For the AR(1) case Δy_{it} is generated as $\Delta y_{it} = \mu_\delta \Delta\delta_t + \gamma\Delta y_{i,t-1} + \Delta\zeta_{it}$, for $i = 1, 2, ..., N; t = 2, ..., T$, with $\Delta y_{i1} = \mu_\delta \Delta\delta_1 + \Delta\zeta_{i1}$ where the process is initalised at $\Delta y_{i0} = 0$, $\gamma = 1$ and $\zeta_{it} = \boldsymbol{\eta}_i'\mathbf{f}_t + u_{it}$. The idiosyncratic errors are generated as $u_{it} \sim IID\mathcal{N}(0, \sigma^2)$ under Gaussianity and $u_{it} \sim IID\frac{\sigma}{\sqrt{12}}(\chi_6^2 - 6)$ under non-Gaussianity where χ_6^2 is a chi-square variate with 6 degrees of freedom, and $\sigma^2 = 1$. The remaining parameters are generated as described in Section 8.1.1. For the ARX(1) case Δy_{it} is generated as $\Delta y_{it} = \mu_\delta \Delta\delta_t + \gamma\Delta y_{i,t-1} + \beta\Delta x_{it} + \Delta\zeta_{it}$, for $i = 1, 2, ..., N; t = 2, ..., T$, with $\Delta y_{i1} = \mu_\delta \Delta\delta_1 + \Delta\zeta_{i1}$ where the process is initalised at $\Delta y_{i0} = 0$, $\gamma = 1$ and $\zeta_{it} = \boldsymbol{\eta}_i'\mathbf{f}_t + u_{it}$. The idiosyncratic errors, u_{it}, are generated as in the AR(1) case with $\sigma^2 = (1 - R^2_{\Delta y})/8R^2_{\Delta y}$ and $R^2_{\Delta y} = 0.4$. The remaining parameters are generated as described in Section 8.1.2. For $m_0 = 0$, ζ_{it} collapses to u_{it} in the above set-ups, and the rest follows accordingly, with $\sigma^2 = (1 - R^2_{\Delta y})/5R^2_{\Delta y}$ in the case of the ARX(1) model. Each f_t is generated once and the same $f_t's$ are used throughout the replications. The first observation is discarded. $\hat{m}$ is the estimated number of factors computed using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$. All experiments are based on 1,000 replications.

S.5.1.1 Performance of the (Q)ML estimator

The next set of results concern the performance of the proposed estimator when the number of factors is estimated based on the sequential MTLR procedure. The results for the case where the number of factors is known are also included for comparison. Results are reported for the case of Gaussian and non-Gaussian errors.

AR(1) Simulation results for the AR(1) model are provided in Tables S12 and S13. These tables report the bias and RMSE, both multiplied by 100, as well as empirical size and power for the QML estimates of γ. The number of factors, when estimated, is computed based on the sequential MTLR procedure described in Section 7.1 with the significance level $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$. The results show that both the bias and RMSE are sufficiently small and the empirical size is close to the nominal level regardless of whether the number of factors is estimated or not, and the error term is Gaussian or not.

Table S12. Bias($\times$100) and RMSE($\times$100) of γ for the AR(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0, ($\gamma = 1$)

N	$T=5$								$T=10$							
	Bias ($\times$100)				RMSE ($\times$100)				Bias ($\times$100)				RMSE ($\times$100)			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
							Gaussian									
100	0.02	-0.54	-2.13	-3.46	3.31	7.08	10.66	15.90	0.01	0.00	0.02	-0.07	1.20	1.47	1.23	2.21
300	0.01	0.02	-1.04	-1.23	2.05	2.05	7.79	8.16	0.02	0.00	0.01	0.00	0.71	0.69	0.70	0.69
500	0.01	0.01	-0.98	-0.91	1.64	1.82	7.22	7.32	0.01	0.00	0.02	-0.01	0.54	0.53	0.55	0.55
							non-Gaussian									
100	-0.04	-0.95	-2.88	-4.71	3.35	8.57	12.30	18.02	-0.01	-0.14	0.05	-0.17	1.19	2.20	1.25	3.13
300	0.03	-0.35	-0.98	-1.99	2.17	4.99	7.64	11.32	-0.02	-0.01	0.02	-0.02	0.70	0.97	0.69	1.05
500	0.07	-0.23	-0.73	-1.37	1.65	4.64	6.64	9.51	-0.01	-0.01	0.03	-0.06	0.50	0.66	0.52	1.50

Note: $\hat{m}$ is estimated using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$; γ is the coefficient of the lagged dependent variable given in (1) in the absence of the $\mathbf{x}_{it}$ regressors. All experiments are based on 2,000 replications. See also the notes to Table S11.

Table S13. Size(%) and power(%) of γ for the AR(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0 ($\gamma = 1$)

N	$T=5$								$T=10$							
	Size				Power				Size				Power			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
								Gaussian								
	H_0: $\gamma = 1$				H_1: $\gamma = 0.96$				H_0: $\gamma = 1$				H_1: $\gamma = 0.96$			
100	3.2	3.7	5.7	6.6	22.6	23.7	23.2	23.2	4.9	4.1	4.7	4.9	90.4	90.3	88.4	88.4
300	5.1	4.8	6.4	6.1	50.6	49.9	44.1	44.5	6.2	5.0	5.3	4.5	100.0	100.0	100.0	100.0
500	5.8	5.6	6.1	5.4	69.9	70.3	62.7	61.3	5.8	5.4	5.7	4.8	100.0	100.0	100.0	100.0
								non-Gaussian								
	H_0: $\gamma = 1$				H_1: $\gamma = 0.96$				H_0: $\gamma = 1$				H_1: $\gamma = 0.96$			
100	3.4	5.1	5.8	7.2	23.5	24.7	22.0	22.9	5.8	5.5	6.2	6.4	90.1	87.5	86.2	85.2
300	5.2	4.7	4.6	6.0	49.6	50.0	40.9	42.6	5.5	4.6	4.7	5.3	100.0	99.7	99.9	99.5
500	4.8	5.2	5.9	5.1	66.5	67.6	55.1	55.5	4.2	5.7	4.2	5.8	100.0	99.9	100.0	99.8

See the notes to Table S12.

ARX(1) Simulation results for the ARX(1) model are provided in Tables S14 and S15. Similar results as in the AR(1) model are found for the ARX(1). The bias and RMSE are small and inference is accurate with reasonably high power regardless of whether the number of factors is estimated or not, and the error term is Gaussian or not.

Table S14. Bias($\times$100) and RMSE($\times$100) of γ and β for the ARX(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0 ($\gamma = 1$)

N	$T=5$								$T=10$							
	Bias ($\times$100)				RMSE ($\times$100)				Bias ($\times$100)				RMSE ($\times$100)			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
								Gaussian								
γ, $\gamma = 1$																
100	0.05	0.01	0.04	0.04	1.55	1.53	1.52	1.47	0.00	0.00	0.01	0.03	0.45	0.47	0.47	0.48
300	0.03	-0.01	-0.02	0.04	0.88	0.89	0.84	0.83	0.01	0.00	0.00	0.01	0.25	0.26	0.26	0.26
500	0.03	-0.02	0.02	-0.03	0.67	0.67	0.63	0.63	0.00	0.01	0.00	0.00	0.20	0.20	0.21	0.21
β																
100	-0.06	-0.12	-0.05	0.09	3.83	3.88	3.87	3.88	-0.07	0.04	-0.04	-0.09	2.18	2.19	2.26	2.30
300	-0.04	0.08	0.02	0.07	2.17	2.20	2.24	2.20	0.00	0.00	0.01	-0.06	1.27	1.22	1.25	1.26
500	0.01	-0.06	-0.03	-0.03	1.68	1.68	1.68	1.71	-0.04	0.00	0.03	0.01	0.95	0.94	0.95	0.98
								non-Gaussian								
γ, $\gamma = 1$																
100	0.06	0.07	0.02	0.03	1.60	1.55	1.47	1.48	-0.03	0.01	-0.01	0.00	0.46	0.47	0.47	0.47
300	-0.02	-0.02	0.04	-0.02	0.88	0.91	0.86	0.84	0.00	0.01	0.01	0.00	0.26	0.26	0.26	0.27
500	0.01	0.00	-0.02	0.01	0.70	0.69	0.65	0.64	0.00	0.01	0.00	0.00	0.20	0.20	0.21	0.20
β																
100	-0.18	0.01	0.07	-0.13	3.83	3.74	3.89	4.02	0.08	-0.03	0.00	0.00	2.14	2.19	2.16	2.21
300	-0.05	0.07	-0.07	0.14	2.12	2.21	2.23	2.18	0.02	0.00	0.01	0.01	1.22	1.27	1.28	1.29
500	0.01	0.01	0.03	-0.01	1.66	1.63	1.71	1.71	0.00	-0.04	-0.01	0.01	0.97	0.97	1.00	0.98

Note: $\hat{m}$ is estimated using the sequential MTLR procedure described in Section 7.1 with $\alpha_N = 50\frac{p}{(T-2)N}$ and $p = 0.05$; γ and β are the coefficients of the lagged dependent variable and the $\mathbf{x}_{it}$ regressor given in (1). All experiments are based on 2,000 replications. See also the notes to Table S11.

Table S15. Size(%) and power(%) of γ and β for the ARX(1) model, using the estimated number of factors, $\hat{m}$, and the true number, m_0 ($\gamma = 1$)

N	$T=5$								$T=10$							
	Size				Power				Size				Power			
(m, m_0)	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$	$(1,1)$	$(\hat{m},1)$	$(2,2)$	$(\hat{m},2)$
Gaussian																
γ	H_0: $\gamma = 1$				H_1: $\gamma = 0.98$				H_0: $\gamma = 1$				H_1: $\gamma = 0.98$			
100	5.2	5.5	6.4	5.4	26.0	26.2	29.7	30.7	4.4	5.7	5.9	6.1	99.3	98.8	98.6	98.6
300	4.9	5.5	6.4	5.4	61.7	63.3	68.2	67.0	3.7	5.2	4.5	5.0	100.0	100.0	99.9	99.9
500	4.7	4.8	5.0	4.8	82.2	84.3	87.2	88.0	5.4	4.6	5.6	6.2	100.0	100.0	100.0	100.0
β	H_0: $\beta = 1$				H_1: $\beta = 0.95$				H_0: $\beta = 1$				H_1: $\beta = 0.95$			
100	5.4	5.5	5.6	5.4	27.3	28.9	25.8	25.0	5.6	5.4	6.2	6.4	65.3	64.1	62.7	64.2
300	5.3	5.5	5.6	5.3	64.5	60.4	61.7	61.4	5.6	4.7	4.8	5.2	97.5	97.5	97.4	97.7
500	4.9	5.6	4.6	5.2	83.9	83.4	84.0	82.7	5.4	5.0	4.1	5.7	99.8	99.7	99.7	99.9
non-Gaussian																
γ	H_0: $\gamma = 1$				H_1: $\gamma = 0.98$				H_0: $\gamma = 1$				H_1: $\gamma = 0.98$			
100	6.6	5.6	6.0	6.0	26.7	25.7	31.1	30.0	6.0	6.9	5.6	6.0	97.8	97.0	97.3	97.4
300	5.0	5.6	5.6	5.4	62.8	61.8	64.5	67.6	4.7	4.4	4.4	4.8	99.1	98.8	99.2	99.1
500	5.9	5.1	5.3	5.0	81.5	79.8	87.2	85.6	4.9	3.9	5.2	4.6	98.9	98.9	99.3	98.9
β	H_0: $\beta = 1$				H_1: $\beta = 0.95$				H_0: $\beta = 1$				H_1: $\beta = 0.95$			
100	6.3	5.1	7.2	6.8	28.8	26.0	25.4	28.4	5.1	5.7	5.1	5.2	61.8	62.1	60.7	59.9
300	4.5	5.7	5.3	4.5	64.4	57.6	62.1	58.7	3.9	4.9	4.5	5.0	93.7	93.3	93.9	92.9
500	5.1	4.1	4.8	5.5	84.5	78.0	82.7	79.4	4.8	4.6	4.5	4.1	97.3	96.1	96.1	95.3

See the notes to Table S14.