ARTICLE IN PRESS
Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

An incidental parameters free inference approach for panels with common shocks[☆]

Artūras Juodis [a,*], Vasilis Sarafidis [b,1]

[a] University of Amsterdam and Tinbergen Institute, The Netherlands
[b] BI Norwegian Business School, Norway

ARTICLE INFO

Article history:
Received 10 June 2019
Received in revised form 4 March 2021
Accepted 19 March 2021
Available online xxxx

JEL classification:
C13
C15
C23

Keywords:
Common factors
GMM
Incidental parameter problem
Endogenous regressors
U-statistic

ABSTRACT

This paper develops a novel Method of Moments approach for panel data models with endogenous regressors and unobserved common factors. The proposed approach does not require estimating explicitly a large number of parameters in either time-series or cross-sectional dimension, T and N respectively. Hence, it is free from the incidental parameter problem. In particular, the proposed approach does not suffer from "Nickell bias" of order $O(T^{-1})$, nor from bias terms that are of order $O(N^{-1})$. Therefore, it can operate under substantially weaker restrictions compared to existing large T procedures. Two alternative GMM estimators are analyzed; one makes use of a fixed number of "averaged estimating equations" à la Anderson and Hsiao (1982), whereas the other one makes use of "stacked estimating equations", the total number of which increases at the rate of $O(T)$. It is demonstrated that both estimators are consistent and asymptotically mixed-normal as $N \to \infty$ for any value of T. Low-level conditions that ensure local and global identification in this setup are examined using several examples.

1. Introduction

Common factor structures are a topic of broad interest in panel data analysis because they offer wide scope for controlling for unobservables, including situations where there is cross-sectional dependence (see e.g. Chudik and Pesaran, 2015b and Sarafidis and Wansbeek, 2021, among others).

To date, the panel common factor literature has been divided among methods catering for panels where the number of time series observations, denoted by T, is "fixed and small", and panels where "T is large". The present paper develops a new Generalized Method of Moments (GMM) approach, which is consistent for any value of T. Currently, there are no results in the GMM literature covering the setup where T can be small or large. Our approach is mainly motivated by the increasing availability of micro-panels in which the value of T is not negligible (but often not as large as in macro-panels),

☆ We would like to thank Serena Ng (the Editor), an anonymous Associate Editor and two anonymous referees for numerous suggestions that greatly improved this paper. We would also like to thank participants at the International Panel Data Conference (Thessaloniki, 2017; Seoul, 2018) and the (EC)^2 Conference (Oxford, 2019) for useful feedback. We also thank Maurice Bun, Frank Kleibergen, Roger Moon and Martin Weidner for constructive comments and advice on the numerous stages of this research project. Financial support from the Netherlands Organization for Scientific Research (NWO) under research grant number 451-17-002 is gratefully acknowledged by the first author.

* Correspondence to: Amsterdam School of Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands.
E-mail addresses: A.Juodis@uva.nl (A. Juodis), vasilis.sarafidis@aims.edu.au (V. Sarafidis).
1 Department of Economics, BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway.

https://doi.org/10.1016/j.jeconom.2021.03.011

Please cite this article as: A. Juodis and V. Sarafidis, An incidental parameters free inference approach for panels with common shocks. Journal of Econometrics (2021), https://doi.org/10.1016/j.jeconom.2021.03.011.

as it is the case with several household surveys,[2] or firm level panels that contain balance sheet and income statement data.[3]

The proposed approach gives rise to estimators that are free from incidental parameters by construction. In particular, two key elements are combined: (i) a quasi-differencing transformation that removes the unobserved *factor loadings* from the error; and (ii) an approximation of the unknown *factors*, based either on observed variables or on the composite error term of the model. The latter has an exact factor structure once evaluated at the true value of the slope parameters. Essentially, these two elements allow us to devise a strategy that avoids estimating explicitly a large number of parameters, regardless of the size of N or T.

More specifically, we put forward two alternative GMM estimators; one is based on a constant number of "averaged estimating equations" à la Anderson and Hsiao (1982), whereas the other one makes use of "stacked estimating equations", the total number of which increases at the rate of $\mathcal{O}(T)$. We demonstrate that the former estimator is consistent and asymptotically mixed-normal as $N \to \infty$ for any value of T. The latter remains consistent and asymptotically mixed-normal although, unsurprisingly, it can be subject to an asymptotic bias proportional to T/N due to the use of "many moment conditions".

For the case of a simple AR(1) model with a one-way error components structure, it is shown that the proposed estimating equations reduce to moment conditions with instruments either in first-differences or long-differences, as e.g. in Hahn et al. (2007) (see Example 1 in Section 3.2 for details).

In comparison to the approach developed in this paper, existing fixed T GMM estimators require estimation of $\mathcal{O}(T)$ nuisance parameters in order to control for the unobserved factors; see e.g. Holtz-Eakin et al. (1988), Nauges and Thomas (2003), Hayakawa (2011), Ahn et al. (2013), Robertson and Sarafidis (2015) and Juodis and Sarafidis (2020), among others. Such requirement is problematic for asymptotic approximations with $T \to \infty$. The reason is that the moment conditions become multiplicative functions of incidental parameters in the T dimension, and thereby standard GMM formulation breaks down.[4] Unfortunately, alternative fixed T consistent estimators proposed in the literature, such as the "correlated random effects" Maximum Likelihood (ML) approach of Bai (2013), are only applicable to the panel AR(1) model with strictly exogenous regressors. As such, they are not suitable for our general setup, which allows for endogeneity and general forms of weak exogeneity.

For panels with T large, popular estimators include those developed by Pesaran (2006) and Bai (2009), known in the literature as "CCE" (Common Correlated Effects) and "PC" (Principal Components), respectively. Both CCE and PC have been originally designed for models with strictly exogenous regressors. Recent extensions to models with weakly exogenous or endogenous regressors suffer from the "incidental parameters problem" since an increasing number of nuisance parameters needs to be estimated as either T or N grows. That is, unbiased asymptotic inference is guaranteed only after appropriate bias-correction; see e.g. Lee et al. (2012), Everaert and Pozzi (2014), Chudik and Pesaran (2015a), Moon and Weidner (2017), Juodis et al. (2021) and Juodis (2020a), among others.[5] By contrast, the GMM approach in this paper does not suffer from "Nickell bias" of order $\mathcal{O}(T^{-1})$, nor from bias terms that are of order $\mathcal{O}(N^{-1})$, where N denotes the number of cross-sectional observations. Therefore, it can operate under substantially weaker restrictions compared to existing large T procedures, which typically require $T \approx N$.

The proposed approach enables us to transparently study identification in this setup. In particular, we investigate several examples and provide necessary and sufficient conditions to ensure local and global identification of the slope parameters. Currently, the vast majority of the GMM panel factor literature takes it for granted that the moment conditions globally identify the parameters of the model. This is despite the fact that global identification might fail for nonlinear moment conditions, which existing GMM procedures heavily rely upon.[6]

One interesting result shows that when it comes to identification of the true parameter vector, one cannot do worse by using stacked estimating equations as opposed to averaged ones. That is, the class of globally and locally identified models based on averaged estimating equations is no larger than the class based on stacked ones. In terms of local identification specifically, we demonstrate that even when the Jacobian matrix of the averaged estimating equations is singular for T large, it converges to zero at a slower rate than the moment conditions. Thus, the corresponding GMM estimator is at least $\sqrt{N}$-consistent and mixed-normal. Notably, inference remains valid without knowledge of the convergence rate of the estimator. That is, from a practical point of view, there is no need to know the convergence rate of the estimator in order to conduct asymptotically valid inferences. For these reasons, this setup is different from the weak-identification setup in Staiger and Stock (1997), where failure of local identification may lead to inconsistent parameter estimates. In

2 Prominent examples include the Panel Study of Income Dynamics (PSID) in the U.S. and the European Union Labor Force Survey (EC LFS), which contains quarterly data spanning the period 1983–present.

3 For example, the Federal Deposit Insurance Corporation (FDIC) provides detailed data for the U.S. banking industry on a quarterly frequency from 2001 onwards.

4 Note that even in the literature of GMM estimation with "many moment conditions", it is typically assumed that the number of parameters is fixed as the number of moment conditions grows large, see e.g. Han and Phillips (2006) and Newey and Windmeijer (2009).

5 A notable exception is the IV estimator of Norkutė et al. (2021) and Cui et al. (2020), which makes use of averaged moment conditions, and requires strictly exogenous regressors as well as $N \approx T$.

6 Hayakawa (2016) provides examples where global identification fails in the case where $T = 3$, for the moment conditions proposed by Ahn et al. (2013).

terms of global identification, we show that potential failure of global identification can be avoided in a straightforward manner, using multiple factor proxies.

Finally, the proposed approach is appealing because it can be extended to a wide range of models, motivated by either the micro- or macro-econometric literature. These include non-parametric models (e.g. Su and Jin, 2012), models with spatial dependence (Kuersteiner and Prucha, 2020), unit root tests (Robertson et al., 2018), smooth transition and structural breaks (Qian and Su, 2016), inference in partially identified panels with common factors (Hong et al., 2019), to mention a few.

The remainder of this paper is as follows: Section 2 presents the model and puts forward estimating equations, which are either linear or nonlinear in the parameters of interest. Section 3 develops limit theory for the proposed GMM estimators under $N, T \to \infty$ asymptotics. Section 4 discusses various extensions, including models with multiple factors. Section 5 studies inference and identification. Section 6 examines the finite sample properties of the proposed estimators using Monte Carlo experiments. A final section concludes. Proofs of theoretical results are documented in Appendix A. An online Supplementary Appendix studies the properties of the proposed GMM estimators under large N, fixed T asymptotics. Furthermore, the Supplementary Appendix explores identification-robust inference and analyzes local and global identification for the panel AR(1) model, both theoretically and numerically.

2. Framework

2.1. The model

We consider the following linear panel data model with a single factor component:

$$y_{i,t} = \boldsymbol{x}_{i,t}'\boldsymbol{\beta} + \lambda_i f_t + \varepsilon_{i,t}; \quad i = 1, \dots, N; \quad t = 1, \dots, T, \tag{1}$$

where $\boldsymbol{x}_{i,t} = \left(x_{i,t}^{(1)}, \dots, x_{i,t}^{(K)}\right)'$ denotes a $[K \times 1]$ vector of regressors, $\boldsymbol{\beta}$ is the corresponding slope parameter vector, whereas λ_i and f_t denote the individual-specific factor loadings and factors, respectively.[7]

Throughout this paper, $\boldsymbol{x}_{i,t}$ is allowed to be correlated both with λ_i and f_t, although $\boldsymbol{x}_{i,t}$ need not have a factor structure and it can be a nonlinear function of λ_i, f_t. Furthermore, the cross-sectional correlation in $\boldsymbol{x}_{i,t}$ is not restricted to be driven by f_t only.[8] To simplify some expressions, in what follows we shall use the shorthand notation $N_1 = N - 1$ and $T_1 = T - 1$.

Suppose for the moment that the sequence of factors $\{f_t\}_{t=1}^T$ is observed. In order to remove the source of endogeneity that stems from the dependence between $\boldsymbol{x}_{i,t}$ and λ_i, we make use of the following *one-step Forward* Quasi-Differencing (hereafter, FQD) transformation:

$$f_{t+1}(y_{i,t} - \boldsymbol{x}_{i,t}'\boldsymbol{\beta}) - f_t(y_{i,t+1} - \boldsymbol{x}_{i,t+1}'\boldsymbol{\beta}) = f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}; \quad t = 1, \dots, T_1, \tag{2}$$

noting that

$$f_{t+1}\lambda_i f_t - f_t\lambda_i f_{t+1} = 0. \tag{3}$$

Eq. (2) resembles the one-step backward quasi-differencing transformation employed by Holtz-Eakin et al. (1988), except that therein the above equation is divided by f_{t+1}, mutatis mutandis. That is, in their case the factor is essentially reparameterized as $r_t = f_t/f_{t+1}$. As a result, consistent parameter estimation requires that $f_{t+1} \neq 0$ for *all* t. By contrast, the above FQD transformation only requires $f_{t+1} \neq 0$ for at least one value of t. Notice that the FQD transformation can be extended to accommodate the Quasi-Long-Differencing (hereafter, QLD) transformation proposed by Ahn et al. (2013), after replacing $t + 1$ with T. However for technical reasons, the QLD transformation is not appealing in the present setup because it effectively conditions on the last value of the factor, i.e. f_T becomes present in all equations. Therefore, the unconditional limiting distribution of the resulting GMM estimator becomes a function of f_T in the limit as $T \to \infty$. While such dependence is irrelevant for T fixed, it is not appropriate (or not even defined) for $T \to \infty$.[9]

In what follows, all random variables are defined on the common probability space $(\Omega, \mathcal{A}, P)$. Furthermore, $\mathcal{F}$ denotes the σ-field generated by all common shocks driving the individual specific variables in the model, such that conditionally on $\mathcal{F}$ all cross-sectional units are independent. In particular, all factors $\{f_t\}_{t=1}^T$ are measurable with respect to $\mathcal{F}$, but we also allow variables such as regressors and instruments to be a function of other common shocks (not necessarily of finite dimension), resulting in additional sources of dependence across cross-sectional units.[10]

Under standard assumptions to be documented later, it holds from Eq. (2) that

$$\mathrm{E}_{\mathcal{F}}\left[f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}\right] = 0, \quad t = 1, \dots, T_1. \tag{4}$$

7 For ease of exposition, we consider a single factor model. The case of multiple factors is discussed in Section 4.

8 In this respect, our framework resembles closer to the framework in Bai (2009) than in Pesaran (2006).

9 One can also use forward differencing that involves higher steps, such as two-step or three-step. However, one-step differencing is attractive in the single factor model because it retains the maximum possible number of observations in estimation.

10 For example, one can easily allow $\boldsymbol{x}_{i,t} = \boldsymbol{b}\left(\boldsymbol{\psi}_i, \boldsymbol{g}_t, \zeta_{i,t}\right)$, where $\boldsymbol{b}(\cdot)$ is a linear/nonlinear function in all arguments. This is similar to models discussed in Menzel (2019), Juodis (2020b) and Fernández-Val et al. (2021).

In order to allow for an additional source of endogeneity stemming from possible correlations between $\boldsymbol{x}_{i,t}$ and the idiosyncratic error component, $\varepsilon_{i,t}$, we assume that there exists a D_t dimensional vector of instruments, $\boldsymbol{z}_{i,t}$, such that

$$\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,s}|\boldsymbol{z}_{i,t}] = 0; \quad s \geq t. \tag{5}$$

$\boldsymbol{z}_{i,t}$ may contain elements of $\boldsymbol{x}_{i,t}$ (or lagged values thereof), depending on whether the regressors are strictly/weakly exogenous or endogenous with respect to $\varepsilon_{i,t}$. As it is the case with $\boldsymbol{x}_{i,t}$, the correlation between $\boldsymbol{z}_{i,t}$ and $\lambda_i f_t$ is left unrestricted. Note that the restriction on conditional moments in Eq. (5) is sufficient but not necessary, as the estimators developed in this paper employ unconditional moments of the form $\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,s}\boldsymbol{z}_{i,t}] = \boldsymbol{0}_{D_t}$, $s \geq t$. However, for technical reasons related to the large T theory, it is more convenient to specify (more restrictive) conditional moment restrictions. For more details see Appendix A.

Given the moment conditions listed in Eqs. (4)–(5), we put forward the following estimating equations indexed by $\boldsymbol{\beta}$, which will be used to develop Method of Moments estimators for $\boldsymbol{\beta}$:

$$\boldsymbol{m}_{i,t}(\boldsymbol{\beta}) = f_{t+1}\boldsymbol{z}_{i,t}(y_{i,t} - \boldsymbol{x}'_{i,t}\boldsymbol{\beta}) - f_t\boldsymbol{z}_{i,t}(y_{i,t+1} - \boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}); \quad t = 1, \ldots, T_1, \tag{6}$$

such that under Eqs. (4)–(5), we have $\mathrm{E}_{\mathcal{F}}\left[\boldsymbol{m}_{i,t}(\boldsymbol{\beta}_0)\right] = \boldsymbol{0}_{D_t}$.

As it stands, the above expression is not feasible because $\{f_t\}_{t=1}^T$ is unobserved. The standard (T "fixed and small") approach, as e.g. in Holtz-Eakin et al. (1988), treats $\{f_t\}_{t=1}^T$ as unrestricted parameters to be estimated. However, this strategy is problematic for asymptotic approximations with $T \to \infty$. This is due to the fact that, unlike the one-way error components model, here the moment conditions become multiplicative functions of incidental parameters, and therefore standard GMM formulation breaks down. Rather than treating the factors as explicit parameters, the present paper puts forward two approaches that circumvent the need to numerically estimate $\{f_t\}_{t=1}^T$ (or $\{\lambda_i\}_{i=1}^N$).

Remark 1 (*Additive Error Components*)**.** The model in Eq. (1) can be extended to control for additional additive error components. For instance, individual-specific effects can be eliminated by taking first-differences a priori. Furthermore, additive time effects can be removed by transforming the model in terms of time-specific cross-sectional averages. The implication of this transformation on the properties of the moment conditions put forward in the present paper is analyzed in Lemma S.1 in the Supplementary Appendix.

Remark 2 (*Low-rank Regressors*)**.** When $\boldsymbol{x}_{i,t}$ includes time-invariant regressors, identification of $\boldsymbol{\beta}$ requires that f_t is sufficiently time-varying, i.e. the factor component does not degenerate. Otherwise, time-invariant regressors are asymptotically eliminated by the quasi-differencing transformation. See Ahn et al. (2001) for further details. On the other hand, the effect of individual-invariant regressors is identifiable, provided that these regressors are not linearly dependent with $\{f_t\}_{t=1}^T$, and one does not transform the model a priori in terms of deviations from time-specific cross-sectional averages, as in Remark 1.

2.2. Linear approach

We start with the simplest possible approach, which requires relatively more restrictions on the data generating process (DGP) but simplifies estimation considerably. In particular, suppose there exists a time-varying variable of the following form:

$$d_{i,t} = \lambda_i^d f_t + \varepsilon_{i,t}^d. \tag{7}$$

Here $d_{i,t}$ can be internal, i.e. one of the regressors in $\boldsymbol{x}_{i,t}$, or external to the model in Eq. (1), as e.g. in Hansen and Liao (2018) among many others.

Remark 3. The existence of $d_{i,t}$ is plausible in panel data models because economic agents are subject to common influences, such as shifts in technology and productivity, changes in preferences and tastes, to mention a few. Therefore, many variables share the same common factors. For example, Pesaran et al. (2013) develop panel unit root tests based on simple averages of cross-sectionally augmented Sargan–Bhargava statistics. In their empirical illustration, they test the null of a unit root in real interest rates across a sample of countries. The unobserved factors are approximated using cross-sectional averages of two external variables, namely oil and equity prices. Juodis and Sarafidis (2020) develop a linear GMM estimator for fixed T panels with unobserved common factors. In their empirical application, they estimate the price elasticity of residential water demand conditional on weather conditions, namely rainfall and temperature. Similar to Pesaran et al. (2013), the unobserved factor component is approximated by an external variable, i.e. the average daily soil moisture index.

Remark 4 (*Comparison with Existing Literature*)**.** Although the use of observed variables to approximate factors draws from existing literature, namely the CCE approach of Pesaran (2006) and the GMM approach of Juodis and Sarafidis (2020), there exist some major differences: firstly, unlike the aforementioned papers, the present approach remains applicable even when $d_{i,t}$ does not exist, as it will be shown below. Secondly, Pesaran (2006) assumes that all K

regressors are strictly exogenous with respect to $\varepsilon_{i,t}$. In contrast, here the regressors can be weakly exogenous or endogenous. Finally, Juodis and Sarafidis (2020) focus solely on the case where T is fixed. In particular, their moment conditions involve nuisance parameters that absorb the (unobserved) correlations between the factor component and the instruments. Since these correlations are allowed to be time-varying, the number of nuisance parameters increases with T, leading to incidental parameters in the time-series dimension. In contrast, the present approach employs a forward quasi-differencing transformation that eliminates $\{\lambda_i\}_{i=1}^N$ from the model; therefore, it avoids numerical estimation of nuisance parameters. However, in comparison to Juodis and Sarafidis (2020), as well as other fixed-T GMM procedures, our approach is expected to be less efficient because we use only $\mathcal{O}(1)$ moment conditions, or at most $\mathcal{O}(T)$, as opposed to $\mathcal{O}(T^2)$ moment conditions. This is illustrated in the context of Example 1, Section 3.2.

Next to Eq. (7) we further assume there exists a time-varying variable, $q_{i,t}$, which is either stochastic or non-stochastic, such that the following condition holds true:

$$\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,s}^d|q_{i,t}] = 0; \quad s \geq t. \tag{8}$$

The above condition implies weak (sequential) exogeneity of $q_{i,t}$ with respect to $\varepsilon_{i,t}^d$. For instance, time-varying weights that may satisfy weak exogeneity are lagged values of $\boldsymbol{z}_{i,t}$. Moreover, let

$$\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d] \neq 0; \quad t = 1, \ldots, T. \tag{9}$$

Multiplying the observed variable $d_{i,s}$ by $q_{i,t}$, for $s = \{t; t+1\}$, one obtains

$$w_{i,t,s}^{(L)} = q_{i,t}d_{i,s} = q_{i,t}\lambda_i^d f_s + q_{i,t}\varepsilon_{i,s}^d, \tag{10}$$

such that

$$w_{t,s}^{(L)} = \mathrm{E}_{\mathcal{F}}\left[w_{i,t,s}^{(L)}\right] = \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d]f_s + \mathrm{E}_{\mathcal{F}}[q_{i,t}\varepsilon_{i,s}^d] = c_t f_s. \tag{11}$$

Hence, $q_{i,t}$ can be thought of as a weight that scales the cross-sectional average of $d_{i,s}$. Using $w_{t,s}^{(L)}$ in place of f_s in Eq. (6), $s = \{t; t+1\}$, we have

$$\boldsymbol{m}_{i,t}^{(L)}(\boldsymbol{\beta}) = w_{t,t+1}^{(L)}\boldsymbol{z}_{i,t}(y_{i,t} - \boldsymbol{x}_{i,t}'\boldsymbol{\beta}) - w_{t,t}^{(L)}\boldsymbol{z}_{i,t}(y_{i,t+1} - \boldsymbol{x}_{i,t+1}'\boldsymbol{\beta}); \quad t = 1, \ldots, T_1, \tag{12}$$

noting that

$$w_{t,t+1}^{(L)}\lambda_i f_t - w_{t,t}^{(L)}\lambda_i f_{t+1} = c_t f_{t+1}\lambda_i f_t - c_t f_t \lambda_i f_{t+1} = 0. \tag{13}$$

Thus $\mathrm{E}_{\mathcal{F}}[\boldsymbol{m}_{i,t}^{(L)}(\boldsymbol{\beta}_0)] = \boldsymbol{0}_{D_t}$. The superscript "$(L)$" emphasizes that the above equations are linear in $\boldsymbol{\beta}$, and therefore they can be used to construct a GMM objective function with a closed form solution; see Section 3.

Remark 5 (*Violation of Weak Exogeneity for $q_{i,t}$*)**.** When $\mathrm{E}_{\mathcal{F}}[q_{i,t}\varepsilon_{i,s}^d] \neq 0$, $w_{t,s}^{(L)}$ has an additional, non-negligible term and therefore Eq. (11) is violated. However, such violation leads to a mis-specified model, which implies that this restriction is testable within the GMM framework based on the usual over-identifying restrictions test statistic.

Remark 6 (*Time-invariant Weights*)**.** Our setup also accommodates naturally setups with time-invariant weights $q_{i,t} = q_i$ $\forall t$, as long as all aforementioned conditions are satisfied. For instance, time-invariant weights are considered in Pesaran (2006), which focuses on $q_i = 1$, and in Juodis and Sarafidis (2020), which considers $q_i \in \{1, y_{i,0}, \boldsymbol{x}_{i,0}'\}$, where $\boldsymbol{x}_{i,0}$ denotes the $K \times 1$ vector of initial conditions of the covariates. Fan and Liao (2020) have recently advocated a similar construction of factor proxies, which involves pre-specified (potentially arbitrary) weights q_i.

2.3. Nonlinear approach

In some setups, the requirement of the existence of additional variables ($d_{i,t}$) with an exact factor structure in terms of f_t can be restrictive. In this case the linear estimating equations described above may not be feasible. An alternative approach for approximating the factors, which does not require the existence of $d_{i,t}$, can be based on the composite error term of the model.

In particular, let $q_{i,t}$ be as above, but now the conditions (8)–(9) are with respect to the model error term $\varepsilon_{i,t}$. That is, we assume

$$\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,s}|q_{i,t}] = 0; \quad s \geq t, \tag{14}$$

or, more generally,

$$\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,t}|\mathcal{F}_{i,t}] = 0; \quad \mathcal{F}_{i,t} = \sigma\left(\{\boldsymbol{z}_{i,\tau}\}_{\tau=1}^t \vee \{q_{i,\tau}\}_{\tau=1}^t\right). \tag{15}$$

This is essentially the same assumption placed on $q_{i,t}$ as in the linear approach, except that the idiosyncratic errors and the factor loadings correspond to the process of $y_{i,t}$ rather than the process of $d_{i,t}$. Thus, using similar notation as with the linear estimator, one may set:

$$w_{i,t,s}^{(NL)} = q_{i,t}d_{i,s}^{(NL)}(\boldsymbol{\beta}); \quad d_{i,s}^{(NL)}(\boldsymbol{\beta}) = y_{i,s} - \boldsymbol{x}_{i,s}'\boldsymbol{\beta}; \quad \lambda_i^d = \lambda_i; \quad \varepsilon_{i,s}^d = \varepsilon_{i,s}. \tag{16}$$

Effectively, Eq. (16) makes use of the fact that the composite error term, once evaluated at the true value $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, has an exact factor structure, i.e. $y_{i,s} - \boldsymbol{x}_{i,s}'\boldsymbol{\beta}_0 = \lambda_i f_s + \varepsilon_{i,s}$.

Defining $w_{t,s}^{(NL)} = \mathrm{E}_{\mathcal{F}}[w_{i,t,s}^{(NL)}]$ and replacing $w_{t,s}^{(L)}$ by $w_{t,s}^{(NL)}$ in Eq. (12), the corresponding estimating equations become nonlinear in $\boldsymbol{\beta}$. The resulting approach is more akin to the framework of Bai (2009) in that it does not require the existence of covariates driven by the same factors as those entering directly into the error term of the process for $y_{i,t}$. On the other hand, as this approach is nonlinear, it is computationally more demanding than the linear one.

3. Asymptotic results

This section studies the case where both N and T grow to infinity jointly. The case where T is fixed with $N \to \infty$ is analyzed in the Supplementary Appendix of this paper.

3.1. Assumptions

Let $\mathcal{K}_i$ be the σ-field generated by all time-invariant, individual-specific random variables for unit i. We denote the σ-field generated by all individual- and time-specific variables by $\mathcal{D}$, i.e. $\mathcal{D} = \sigma(\mathcal{F} \vee \{\mathcal{K}_i\}_{i=1}^N)$. Let Ξ be some finite constant independent of N and T. Moreover, let $\boldsymbol{X}_i = \left(\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,T}\right)'$, $\boldsymbol{Z}_i = \left(\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,T}\right)'$, $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i,1}, \ldots, \varepsilon_{i,T}\right)'$, $\boldsymbol{\varepsilon}_i^d = \left(\varepsilon_{i,1}^d, \ldots, \varepsilon_{i,T}^d\right)'$, $\boldsymbol{q}_i = \left(q_{i,1}, \ldots, q_{i,T}\right)'$. Our assumptions directly accommodate both time-invariant (q_i) and time-varying ($q_{i,t}$) weights.

Assumption 3.1 (*Data Generating Process*). The DGP for all i and t satisfies the following restrictions for some $r \geq 4$ and $\delta > 0$:

(a) $\boldsymbol{\Upsilon}_i = (\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i^d, \boldsymbol{q}_i)$ are identically distributed and independent across i, conditional on $\mathcal{F}$. $\boldsymbol{\Upsilon}_i$ are independent across i, conditional on $\mathcal{D}$.

(b) $\boldsymbol{p}_{i,t} = (\boldsymbol{x}_{i,t}', \boldsymbol{z}_{i,t}', \varepsilon_{i,t}, \varepsilon_{i,t}^d, q_{i,t})' \otimes (1, f_t, f_{t+1})'$ is a ($\mathcal{D}$ conditional) α-mixing (or strong-mixing) sequence in t, with mixing coefficients $\alpha_i(m)$ that are measurable w.r.t. $\mathcal{D}$ and satisfy $\sup_i(\alpha_i(m)) = \mathcal{O}(m^{-\mu})$ as $m \to \infty$, with $\mu = 3(r+\delta)/\delta$.[11] Each element $p_{i,t}^{(h)}$ satisfies $\mathrm{E}_{\mathcal{D}}\left[\left|p_{i,t}^{(h)}\right|^{r+\delta}\right] < \Xi$.

(c) σ-fields $\mathcal{K}_i$ are independent across i, conditional on $\mathcal{F}_c \subset \mathcal{F}$. $\boldsymbol{v}_i = (\lambda_i, \lambda_i^d, q_i)'$ is $\mathcal{K}_i$-measurable and identically distributed across i. Furthermore, each element $v_i^{(h)}$ satisfies $\mathrm{E}_{\mathcal{F}}\left[\left|v_i^{(h)}\right|^{r+\delta}\right] < \Xi$.

(d) $\mathrm{E}_{\mathcal{D}}[\varepsilon_{i,t}|\mathcal{F}_{i,t}] = 0$, and $\mathrm{E}_{\mathcal{D}}[\varepsilon_{i,t}^d|\mathcal{F}_{i,t}^d] = 0$, where $\mathcal{F}_{i,t} = \sigma(\{\boldsymbol{z}_{i,\tau}\}_{\tau=1}^t)$ and $\mathcal{F}_{i,t}^d = \sigma(\{q_{i,\tau}\}_{\tau=1}^t)$.

The notion of conditional mixing has been used by Fernández-Val and Weidner (2016), Su et al. (2015) and Lu and Su (2016), among others, in the context of large T panel data analysis. Note that unlike many other studies, e.g. Su et al. (2015), here "innovations" $\varepsilon_{i,t}$, and/or $\varepsilon_{i,t}^d$ are not assumed to be martingale difference sequences (MDS). In particular, the large N dimension allows both random sequences to be serially and contemporaneously correlated, as long as they are mixing. The rates r and μ are generally sufficient to use most of the standard inequalities for mixing processes, see e.g. Doukhan (1994). The conditional mixing restrictions on the memory of $\boldsymbol{p}_{i,t}$ are imposed to ensure that the resulting convergence rate is $\sqrt{NT}$ and not $\sqrt{N}$. However, as discussed by Hansen (2007), the slower convergence rate generally has no impact on standardized statistics in this setup. Alternatively, one can impose high-level assumptions directly on the convergence rates of averages of data, as in Bai (2009). In this way, part (b) of Assumption 3.1 on the $\alpha(m)$-mixing coefficient $\alpha_i(m)$, can be relaxed. Finally, for the nonlinear estimating equations, condition (d) should be interpreted in terms of the joint σ-field $\sigma(\{\boldsymbol{z}_{i,\tau}\}_{\tau=1}^t \vee \{q_{i,\tau}\}_{\tau=1}^t)$.

3.2. Estimating equations

Consider the following $[D_t \times 1]$ vector of estimating equations available at time period t:

$$\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}) = \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \left[w_{j,t,t+1}^{(\xi)}\boldsymbol{z}_{i,t}\left(y_{i,t} - \boldsymbol{x}_{i,t}'\boldsymbol{\beta}\right) - w_{j,t,t}^{(\xi)}\boldsymbol{z}_{i,t}\left(y_{i,t+1} - \boldsymbol{x}_{i,t+1}'\boldsymbol{\beta}\right)\right], \tag{17}$$

for $t = 1, \ldots, T_1$, where $\xi \in \{L; NL\}$. The double summation over (i,j) is a direct by-product of making use of cross-sectional averages of $w_{i,t,s}^{(\xi)}$ in the approximation of f_s, $s = \{t; t+1\}$.

[11] The mixing coefficients are defined as $\alpha_i(m) = \sup_t \sup_{A\in\mathcal{A}_t^i, B\in\mathcal{B}_{t+m}^i} |\mathrm{P}_{\mathcal{D}}(A\cap B) - \mathrm{P}_{\mathcal{D}}(A)\mathrm{P}_{\mathcal{D}}(B)|$, where $\mathcal{A}_t^i$ and $\mathcal{B}_t^i$ denote the σ-field generated by $\left(\boldsymbol{p}_{i,t}, \boldsymbol{p}_{i,t-1}, \ldots\right)$ and $\left(\boldsymbol{p}_{i,t}, \boldsymbol{p}_{i,t+1}, \ldots\right)$, respectively. Intuitively, a stochastic process is mixing if its values at widely-separated times are asymptotically independent. Thus, the mixing coefficients $\alpha_i(m)$ represent a "measure of dependence".

Before we provide formal asymptotic analysis, it is useful to informally characterize the asymptotic properties of the leading component in Eq. (17). In particular, under the regularity conditions listed in Assumption 3.1, the leading term in the asymptotic expansion of $\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0)$ is given by[12]:

$$\sqrt{N}\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\mu}_{i,t}^{(\xi)} + o_P(1), \tag{18}$$

such that

$$\boldsymbol{\mu}_{i,t}^{(\xi)} = \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^{(\xi)}]\boldsymbol{z}_{i,t}\left(f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}\right) - \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]q_{i,t}\left(f_{t+1}\varepsilon_{i,t}^{(\xi)} - f_t\varepsilon_{i,t+1}^{(\xi)}\right), \tag{19}$$

where for convenience we adopt the notation $\lambda_i^{(L)} = \lambda_i^d$, $\lambda_i^{(NL)} = \lambda_i$, $\varepsilon_{i,t}^{(L)} = \varepsilon_{i,t}^d$ and $\varepsilon_{i,t}^{(NL)} = \varepsilon_{i,t}$. Thus, although the original sample estimating equations involve double summation over (i, j), the leading term in the expression above involves single summation over the cross-sectional dimension. This is because averages over j implicitly estimate expected values, i.e. $\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^{(\xi)}]$ and $\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]$.[13]

For the case of the nonlinear approach specifically, it is straightforward to see that Eq. (19) simplifies considerably:

$$\boldsymbol{\mu}_{i,t}^{(NL)} = \left(\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i]\boldsymbol{z}_{i,t} - \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]q_{i,t}\right)\left(f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}\right). \tag{20}$$

The above expression indicates that the proposed method quasi-differences not only the model for $y_{i,t}$ in Eq. (1), to avoid estimation of λ_i, but also functions involving the instruments, $\boldsymbol{z}_{i,t}$. This is due to the fact that the correlation between instruments and the factor component is left unrestricted. By contrast, when $\{f_t\}_{t=1}^T$ is treated as known (e.g. $f_t = 1$ $\forall t$), then one requires differencing of $y_{i,t}$ only, but not $\boldsymbol{z}_{i,t}$. However, it is worth noting that these transformations are implemented without any need to numerically estimate the nuisance parameters $\{\lambda_i\}_{i=1}^N$ and $\{f_t\}_{t=1}^T$.

Example 1. Consider a panel AR(1) model with a one-way error components structure, i.e $f_t = 1$ $\forall t$. Setting $q_{i,t} = y_{i,t-2}$ and $z_{i,t} = y_{i,t-1}$, Eq. (20) reduces to

$$\boldsymbol{\mu}_{i,t}^{(NL)} = \left(\mathrm{E}_{\mathcal{F}}[\lambda_i y_{i,t-2}]y_{i,t-1} - \mathrm{E}_{\mathcal{F}}[\lambda_i y_{i,t-1}]y_{i,t-2}\right)\left(\varepsilon_{i,t} - \varepsilon_{i,t+1}\right). \tag{21}$$

Under mean-stationarity, i.e. $\mathrm{E}_{\mathcal{F}}[\lambda_i y_{i,s}] = (1-\alpha_0)^{-1}\mathrm{E}_{\mathcal{F}}[\lambda_i^2]$, where α_0 denotes the true value of the autoregressive parameter, the above expression simplifies further to

$$\boldsymbol{\mu}_{i,t}^{(NL)} = -\mathrm{E}_{\mathcal{F}}[\lambda_i y_{i,0}]\left(\Delta y_{i,t-2}\Delta\varepsilon_{i,t}\right). \tag{22}$$

That is, in this case the estimating equations reduce to moment conditions with instruments in first-differences (up to a constant), as e.g. in Anderson and Hsiao (1982). On the other hand, with time-invariant weights, $q_{i,t} = q_i = y_{i,0}$ (say), the above expression becomes

$$\boldsymbol{\mu}_{i,t}^{(NL)} = -\mathrm{E}_{\mathcal{F}}[\lambda_i y_{i,0}]\left(y_{i,t-1} - y_{i,0}\right)\Delta\varepsilon_{i,t}, \tag{23}$$

in which case the moment conditions make use of instruments in long-differences. Thus, some of the classical Method of Moments procedures can be viewed as special cases of the estimating equations put forward in the present paper (under some restrictions on the DGP). □

Note that in dynamic panels, moment conditions with instruments in first-differences are known to have larger variance compared to their level counterparts, see e.g. Arellano (1989). Hence, the above example illustrates that the implicit double differencing employed in this paper, i.e. over i and over t, cannot be optimal when any knowledge of either λ_i or f_t is available.

Remark 7 (*Multiple Weights*). The statistical framework considered thus far makes use of a single weight to proxy the factors, which can be time-varying or time-invariant. In essence, this setup corresponds to the exactly identified instrumental variable framework in proxying f_t. Given this natural interpretation, our approach can be easily extended to multiple vector weights $\boldsymbol{q}_{i,t} = \left(q_{i,t}^{(1)}, \ldots, q_{i,t}^{(S)}\right)'$. For instance, for $S = 2$ the estimating equations can be expressed as in the following $[2D_t \times 1]$ vector:

$$\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}) = \begin{pmatrix}\overline{\boldsymbol{m}}_t^{(\xi)(1)}(\boldsymbol{\beta}) \\ \overline{\boldsymbol{m}}_t^{(\xi)(2)}(\boldsymbol{\beta})\end{pmatrix}, \tag{24}$$

where $\overline{\boldsymbol{m}}_t^{(\xi)(\kappa)}(\boldsymbol{\beta})$ corresponds to setting $w_{i,t,s}^{(\xi)(\kappa)} = q_{i,t}^{(\kappa)}d_{i,s}^{(\xi)}$, for $\kappa = 1, 2$. Similarly, for the linear approach one can also consider multiple observed variables, $\boldsymbol{d}_{i,t}$. It is clear from the above formulation in (24) that in practice one does not need

[12] The expression in Eq. (18) can be interpreted as the Hájek projection of $\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0)$, see e.g. Ch. 12 of van der Vaart (2000) for a formal definition.

[13] Essentially these terms are the so-called "g" parameters introduced by Robertson and Sarafidis (2015) and Juodis and Sarafidis (2020). However, unlike these papers, here we do not need to numerically estimate $\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^{(\xi)}]$ and $\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]$.

to choose explicitly among different choices of $\boldsymbol{q}_{i,t}$ or $\boldsymbol{d}_{i,t}$, since these give rise to different estimating equations that can be stacked together, as in a standard overidentifying instrumental variables framework. In this sense, the treatment of multiple weights at time period t, $\boldsymbol{q}_{i,t}$, is no different than the treatment of multiple instruments, $\boldsymbol{z}_{i,t}$. This is in stark difference with the linear GMM estimator of Juodis and Sarafidis (2020), designed for fixed T panels. In particular, therein the correlations between instruments and the factor component (the "g" parameters) are estimated explicitly. Therefore in their setup, including more weights (or variables) than necessary in the approximation of the factors can render the asymptotic distribution of GMM non-standard.[14]

3.3. Limit theory for averaged estimating equations

In what follows, we focus on the case where the dimension of $\boldsymbol{z}_{i,t}$ is fixed for all values of t, such that $D_t = D(= \mathcal{O}(1))$, and we shall study "averaged estimating equations". In particular, we consider the following $[D \times 1]$ vector of estimating equations:

$$\overline{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}) = \frac{1}{T_1}\sum_{t=1}^{T_1} \overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}). \tag{25}$$

In Section 3.4 we shall consider the case where instead of averaging $\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta})$ over t, these estimating equations are stacked such that the total number of moment conditions used in estimation is of order $\mathcal{O}(T)$.

Before stating the remaining assumptions necessary for identification and derivation of the asymptotic distribution of the proposed estimator, the following lemma demonstrates that the linear averaged estimating equations can be biased for T large.

Lemma 1. *Suppose that Assumption 3.1 is satisfied. Then for all $t = 1, \ldots, T_1$,*

$$\mathrm{E}_{\mathcal{D}}\left[\overline{\boldsymbol{m}}_t^{(L)}(\boldsymbol{\beta}_0)\right] \neq \mathbf{0}_D; \tag{26}$$

$$\mathrm{E}_{\mathcal{D}}\left[\overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta}_0)\right] = \mathbf{0}_D. \tag{27}$$

Furthermore, $\boldsymbol{b} = \mathrm{E}\left[\overline{\overline{\boldsymbol{m}}}^{(L)}(\boldsymbol{\beta}_0)\right] = \mathcal{O}(N^{-1})$.

Proof. See Appendix A. □

The form of the (potential) bias for the linear estimating equations is solely determined by the "own" terms and arises because the covariance matrix between $(q_{i,t}, \boldsymbol{z}'_{i,t})'$ and $(\varepsilon_{i,t}, \varepsilon^d_{i,t})'$ is largely unrestricted.[15] Notably, this bias diminishes with large N. To illustrate, suppose that $N = T\rho$ for some positive integer ρ, and the underlying mixing process is strictly stationary. Then, it is straightforward to see that

$$\sqrt{NT}\,\mathrm{E}\left[\overline{\overline{\boldsymbol{m}}}^{(L)}(\boldsymbol{\beta}_0)\right] = \sqrt{\rho}\boldsymbol{b} + o_P(1) = \mathcal{O}_P(1). \tag{28}$$

Thus, the linear estimator might suffer from an "incidental parameters problem" under diagonal (or proportional) asymptotics, which is due to the approximation of the T-dimensional parameter vector $\boldsymbol{f} = (f_1, \ldots, f_T)'$ from NT observations. The bias term is of order $\mathcal{O}(\sqrt{TN^{-1}})$, i.e. it is negligible for $N \gg T$, e.g. when T is fixed. Further details are provided in Appendix A.

Despite of this, the source of the potential bias above is easily eliminated by replacing the factor proxies $\widehat{f}_t = N^{-1}\sum_{i=1}^N w_{i,t,t}^{(L)}$ (identical for all i) with individual-specific proxies:

$$\widehat{f}_{i,t} = \frac{1}{N_1}\sum_{j\neq i}^{N} w_{j,t,t}^{(L)}. \tag{29}$$

The resulting linear estimating equations available for each period t are of the following form:

$$\widetilde{\boldsymbol{m}}_t^{(L)}(\boldsymbol{\beta}) = \frac{1}{N(N_1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left[w_{j,t,t+1}^{(L)}\boldsymbol{z}_{i,t}\left(y_{i,t} - \boldsymbol{x}'_{i,t}\boldsymbol{\beta}\right) - w_{j,t,t}^{(L)}\boldsymbol{z}_{i,t}\left(y_{i,t+1} - \boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}\right)\right]. \tag{30}$$

The corresponding nonlinear equations available for each t are of identical form, except that $w_{j,t,s}^{(L)}$ is replaced by $w_{j,t,s}^{(NL)}$, for $s = \{t; t+1\}$.[16]

14 This issue is circumvented in their paper using regularization or best-subset selection.

15 For the precise definition of the bias term, the interested reader may refer to the corresponding proof.

16 For the nonlinear approach, however, it is not necessary to use individual-specific factor proxies because we know from Lemma 1 that the estimating equations remain unbiased.

The "delete-one" construction of $\widehat{f}_{i,t}$ ensures that factor proxies are uncorrelated with all i specific variables. Therefore, the resulting estimating equations are unbiased for any values of N and T and so they can be viewed as an average *U-statistic* of second degree:

$$\widetilde{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}) = \frac{1}{2}\frac{1}{\binom{N}{2}}\sum_{i=2}^{N}\widetilde{\boldsymbol{m}}_{i,t}^{(\xi)}(\boldsymbol{\beta}), \tag{31}$$

where

$$\begin{aligned}\widetilde{\boldsymbol{m}}_{i,t}^{(\xi)}(\boldsymbol{\beta}) = &\sum_{j<i}\left[w_{j,t,t+1}^{(\xi)}\boldsymbol{z}_{i,t}\left(y_{i,t}-\boldsymbol{\beta}'\boldsymbol{x}_{i,t}\right) - w_{j,t,t}^{(\xi)}\boldsymbol{z}_{i,t}\left(y_{i,t+1}-\boldsymbol{\beta}'\boldsymbol{x}_{i,t+1}\right)\right]\\ &+\sum_{j<i}\left[w_{i,t,t+1}^{(\xi)}\boldsymbol{z}_{j,t}\left(y_{j,t}-\boldsymbol{\beta}'\boldsymbol{x}_{j,t}\right) - w_{i,t,t}^{(\xi)}\boldsymbol{z}_{j,t}\left(y_{j,t+1}-\boldsymbol{\beta}'\boldsymbol{x}_{j,t+1}\right)\right].\end{aligned} \tag{32}$$

Let

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}) = \frac{1}{T_1}\sum_{t=1}^{T_1}\widetilde{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}).$$

For both linear and nonlinear approaches, we define the estimator that makes use of averaged estimating equations as the solution of the following standard Method of Moments minimization problem:

$$\widehat{\boldsymbol{\beta}}_{MM}^{(\xi)} = \arg\min_{\boldsymbol{\beta}\in\boldsymbol{\Theta}}\left(\left(\widetilde{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta})\right)'\boldsymbol{W}_{N,T}\widetilde{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta})\right), \tag{33}$$

for $\xi \in \{L; NL\}$, where $\boldsymbol{W}_{N,T}$ is a positive definite weighting matrix such that $\boldsymbol{W}_{N,T} \xrightarrow{p} \boldsymbol{W}$ as $N, T \to \infty$, and $\boldsymbol{W}$ is assumed to be $\mathcal{F}$-measurable and positive definite a.s. A similar U-statistic based objective function has been recently used in Jochmans (2017) to estimate common parameters for nonlinear dyadic models. As with the setup in Eq. (25) for $\xi = NL$, the proposed moment conditions in Jochmans (2017) are multiplicative functions of common parameters.

The at-most-quadratic nature of the proposed estimating equations implies that

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}) = \widetilde{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}_0) + \left(\widetilde{\overline{\boldsymbol{\Gamma}}}^{(\xi)} + \frac{1}{2}\sum_{k=1}^{K}\widetilde{\overline{\boldsymbol{H}}}_k^{(\xi)}(\beta_k-\beta_{0,k})\right)(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \tag{34}$$

where $\widetilde{\overline{\boldsymbol{\Gamma}}}^{(\xi)} = \left[\partial\widetilde{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is of dimension $[D\times K]$, while the $[D\times K]$ matrices $\widetilde{\overline{\boldsymbol{H}}}_k^{(\xi)}$ denote the corresponding matrix-valued second derivatives of $\widetilde{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, where β_k is the kth element of $\boldsymbol{\beta}$.

Assumption 3.2 (*Local Identification*)**.** For each $\xi = \{L; NL\}$ the limiting Jacobian matrix $\boldsymbol{\Gamma}^{(\xi)} = \text{plim}_{N,T\to\infty}\widetilde{\overline{\boldsymbol{\Gamma}}}^{(\xi)}$ is $\mathcal{F}$-measurable with rank K a.s., i.e. $\text{rk}[\boldsymbol{\Gamma}^{(\xi)}] = K$.

Assumption 3.2 ensures consistency of the proposed estimator based on the linear estimating equations. On the other hand, as it is generally the case for nonlinear approaches, additional restrictions are required to ensure consistency of the proposed nonlinear estimator.

Assumption 3.3 (*Global Identification*)**.** The parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^K$ is compact and contains $\boldsymbol{\beta}_0$ in its interior. The limiting matrices $\text{plim}_{N,T\to\infty}\widetilde{\overline{\boldsymbol{H}}}_k^{(\xi)}$ are $\mathcal{F}$-measurable, and bounded a.s. for all k. Let $\boldsymbol{m}^{(NL)}(\boldsymbol{\beta}) = \text{plim}_{N,T\to\infty}\widetilde{\overline{\boldsymbol{m}}}^{(NL)}(\boldsymbol{\beta})$ for all $\boldsymbol{\beta}\in\boldsymbol{\Theta}$. $\boldsymbol{\beta}_0$ is identified on $\boldsymbol{\Theta}$ such that $\boldsymbol{m}^{(NL)}(\boldsymbol{\beta}) = \boldsymbol{0}_D$ iff $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ a.s.

Assumption 3.3 implies that the corresponding limiting estimating equations point identify the parameter of interest over $\boldsymbol{\Theta}$. In Section 5 we discuss examples where Assumption 3.2–3.3 can be violated.

Denote by $\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}$ the time-series average of the leading term given in Eq. (19), i.e.

$$\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)} = \frac{1}{T_1}\sum_{t=1}^{T_1}\boldsymbol{\mu}_{i,t}^{(\xi)}. \tag{35}$$

We assume that the following quantities are well defined and a.s. finite:

$$\boldsymbol{\Omega}^{(\xi)} = \text{plim}_{T\to\infty}\,\mathrm{E}_{\mathcal{F}}\left[T_1\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}\left(\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}\right)'\right]; \tag{36}$$

$$\boldsymbol{\Sigma}^{(\xi)} = \left[\left(\boldsymbol{\Gamma}^{(\xi)}\right)'\boldsymbol{W}\boldsymbol{\Gamma}^{(\xi)}\right]^{-1}\left(\boldsymbol{\Gamma}^{(\xi)}\right)'\boldsymbol{W}. \tag{37}$$

Moreover, $\boldsymbol{\Sigma}^{(\xi)}$ has rank K a.s. For technical reasons we also impose the following restriction on the relative rates of N, T.

Assumption 3.4 (*Asymptotics*)**.** N_T is a non-decreasing function of T such that $N_T \to \infty$ as $T \to \infty$.

Assumption 3.4 merely requires that N_T is a non-decreasing function of T. In particular, it is weaker than assuming that $T/N \to \rho \in [0;\infty)$, e.g. one can allow $N_T = \sqrt{T}$.[17] This restriction allows us to use the central limit theorem for MDS arrays from Hall and Heyde (1980).

The following theorem summarizes the asymptotic distribution of the proposed estimator.

Theorem 1. *Suppose that Assumptions 3.1–3.4 hold true and* $\mathrm{rk}\left[\boldsymbol{\Omega}^{(\xi)}\right] = D$ *a.s. Then, as* $N, T \to \infty$ *we have*

$$\sqrt{NT_1}\left(\widehat{\boldsymbol{\beta}}_{MM}^{(\xi)} - \boldsymbol{\beta}_0\right) \Rightarrow \boldsymbol{\Sigma}^{(\xi)}\left(\boldsymbol{\Omega}^{(\xi)}\right)^{1/2}\boldsymbol{\psi} \quad (\mathcal{F} - \textit{stably}), \tag{38}$$

for $\xi = \{L; NL\}$, *where* $\boldsymbol{\Sigma}^{(\xi)}$ *and* $\boldsymbol{\Omega}^{(\xi)}$ *are independent of* $\boldsymbol{\psi} \sim N(\mathbf{0}_D, \boldsymbol{I}_D)$.

Proof. See Appendix A. □

Since $\boldsymbol{\Sigma}^{(\xi)}$ and $\boldsymbol{\Omega}^{(\xi)}$ can be random matrices (measurable with respect to $\mathcal{F}$, but independent of $\boldsymbol{\psi}$), the unconditional limiting distribution of the proposed estimator is mixed-normal rather than normal, in general. To appreciate this fact recall that the leading term of the asymptotic distribution is determined by the scaled time-series average in Eq. (35), i.e.

$$\sqrt{T_1}\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)} = \frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\left[\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^{(\xi)}]\boldsymbol{z}_{i,t}\left(f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}\right) - \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]q_{i,t}\left(f_{t+1}\varepsilon_{i,t}^{(\xi)} - f_t\varepsilon_{i,t+1}^{(\xi)}\right)\right]. \tag{39}$$

As $T \to \infty$, all time-varying components average out unless some of them are $\mathcal{F}_c$ measurable. As an example, let $\lambda_i = \lambda_i^d = \lambda\widetilde{\lambda}_i$ where λ is some random variable, and $\widetilde{\lambda}_i$ is an i.i.d. sequence. In such a case, the asymptotic distribution of the estimator (through $\boldsymbol{\Omega}^{(\xi)}$ and $\boldsymbol{\Sigma}^{(\xi)}$) is a function of λ. However, this plays no role for inference procedures that make use of standardized (pivotal) statistics, so long as both $\boldsymbol{\Sigma}^{(\xi)}$ and $\boldsymbol{\Omega}^{(\xi)}$ can be consistently estimated from their sample analogues. That is,

$$NT_1\left(\widehat{\boldsymbol{\beta}}_{MM}^{(\xi)} - \boldsymbol{\beta}_0\right)'\left[\boldsymbol{\Sigma}^{(\xi)}\boldsymbol{\Omega}^{(\xi)}\left(\boldsymbol{\Sigma}^{(\xi)}\right)'\right]^{-1}\left(\widehat{\boldsymbol{\beta}}_{MM}^{(\xi)} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \chi_K^2. \tag{40}$$

The result of Theorem 1 indicates that $\widehat{\boldsymbol{\beta}}_{MM}^{(\xi)}$ is the only available estimator in the literature that does not suffer from incidental parameter bias in any dimension (with the "delete-one" implementation, where required). In particular, our approach does not suffer from bias of order $\mathcal{O}(T^{-1})$ (the so-called "Nickell bias"), because it is based on the Method of Moments with quasi-differenced moment conditions. In contrast, Nickell bias is typical in least-squares type estimators for models with weakly exogenous regressors. Moreover, our approach does not suffer from bias of order $\mathcal{O}(N^{-1})$, because the U-statistic formulation that we employ allows us to avoid numerical estimation of $\{f_t\}_{t=1}^T$. For this reason no explicit restrictions on the relative diagonal expansion rates of N and T are imposed. In comparison, popular large T procedures accommodating a factor structure, such as the so-called PC and CCE estimators, not only have bias terms that are of order $\mathcal{O}(T^{-1})$, but they are also subject to bias terms of order $\mathcal{O}(N^{-1})$; see Moon and Weidner (2017), Juodis et al. (2021), and Juodis (2020a). On the other hand, the quasi-differencing transformation embedded in our approach implies extra computational complexity relative to both PC and CCE.

3.4. Limit theory for stacked estimating equations

In the fixed-T literature of panels with common factors, existing GMM estimators employ stacked moment conditions, such that the total number of instruments used in estimation increases with T, see e.g. Ahn et al. (2013), Robertson and Sarafidis (2015) and Juodis and Sarafidis (2020), among others. This strategy is essential because the number of parameters to be estimated is of order $O(T)$.[18] To the best of our knowledge, there are no theoretical results available in this literature that allow $T \to \infty$.

In what follows we use our approach to study the setup where the number of estimating equations increases with the sample size, in particular, at the rate of $O(T)$.

For both linear and nonlinear approaches, we define the estimator that makes use of stacked estimating equations as the solution of the following minimization problem:

$$\widehat{\boldsymbol{\beta}}_{MMT}^{(\xi)} = \arg\min_{\boldsymbol{\beta}\in\Theta}\left(\frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta})\right)'\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta})\right), \tag{41}$$

[17] Having said that, in most applications we have in mind, $N_T > T$. Note that for simplicity, we drop the T subscript from N_T and simply use N hereafter.

[18] Therefore, averaging of moment conditions is not feasible in their framework.

for $\xi \in \{L; NL\}$ and $\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta})$ as defined in Eq. (17). Here we use the $\widehat{\boldsymbol{\beta}}_{MMT}^{(\xi)}$ notation to emphasize that the number of estimating equations employed is of order $\mathcal{O}(T)$. Moreover, unlike Theorem 1 we do not use the "delete-one" construction in $\widehat{f}_{i,t}$ because (as we show later on) the "many-instruments bias" that arises, is in general of the same order as that of the incidental parameters bias associated with the approximation of f_t from NT observations.

Remark 8 (*Weighting of Stacked Estimating Equations*)**.** In line with existing literature of GMM estimation with many moment conditions (see e.g. Han and Phillips, 2006 and Newey and Windmeijer, 2009), the above estimator is unweighted, i.e. it corresponds to optimizing an objective function with an identity matrix as a "weighting" matrix. There are several reasons for this choice. To begin with, consistent estimation of the optimal weighting matrix can be practically infeasible due to incidental parameters. In particular, while the optimal weighting matrix is typically easy to compute when T is fixed, this is not the case when T is large, as one needs to take the inverse of the variance matrix of the moment conditions, which is of order $\mathcal{O}(T)$ in the present setup. Thus, as it is well appreciated in the literature on high dimensional covariance matrices, the ratio T/N plays a major role for consistent estimation of the corresponding covariance (and precision) matrices in the large T case.[19] Secondly, high dimensional covariance matrices can often be subject to singularities. Unfortunately, the use of a generalized inverse will not solve the problem in the GMM framework.[20] Finally, we note that in principle it is possible to come up with alternative, suboptimal choices for the weighting matrix, the structure of which depends on unknown parameters of fixed dimension. However, such strategy has impact for local and global identification, as these depend explicitly on a particular choice of the weighting matrix.

Analogously to Eq. (34) we expand the estimating equations as follows:

$$\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}) = \overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0) + \left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)} + \frac{1}{2}\sum_{k=1}^{K} \overline{\boldsymbol{H}}_{t,k}^{(\xi)}(\beta_k - \beta_{0,k})\right)(\boldsymbol{\beta} - \boldsymbol{\beta}_0), \tag{42}$$

where all matrices are defined accordingly. Since we consider an increasing number of estimating equations, all regularity conditions imposed in Section 3.3 need to be appropriately modified to accommodate this setup.

Define $\overline{\boldsymbol{\gamma}}_t^{(\xi)} = \left(\mathrm{vec}(\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0))', \mathrm{vec}(\overline{\boldsymbol{\Gamma}}_t^{(\xi)})', \mathrm{vec}(\overline{\boldsymbol{H}}_{t,1}^{(\xi)})', \ldots, \mathrm{vec}(\overline{\boldsymbol{H}}_{t,K}^{(\xi)})'\right)'$.

Assumption 3.5 (*Local Identification: Stacked*)**.** For each $\xi = \{L; NL\}$, $\boldsymbol{\Gamma}_{MMT}^{(\xi)} = \mathrm{plim}_{N,T\to\infty} T_1^{-1} \sum_{t=1}^{T_1} \left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}\right)' \overline{\boldsymbol{\Gamma}}_t^{(\xi)}$ is $\mathcal{F}$-measurable with $\mathrm{rk}[\boldsymbol{\Gamma}_{MMT}^{(\xi)}] = K$ a.s.

Assumption 3.6 (*Global Identification: Stacked*)**.** The parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^K$ is compact and contains $\boldsymbol{\beta}_0$ in its interior. $\mathrm{plim}_{N,T\to\infty} T_1^{-1} \sum_{t=1}^{T_1} \overline{\boldsymbol{\gamma}}_t^{(\xi)} \left(\overline{\boldsymbol{\gamma}}_t^{(\xi)}\right)'$ is $\mathcal{F}$-measurable, and bounded a.s. Let $G^{(NL)}(\boldsymbol{\beta}) = \mathrm{plim}_{N,T\to\infty} T_1^{-1} \sum_{t=1}^{T_1} \overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta})' \overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in \boldsymbol{\Theta}$. $\boldsymbol{\beta}_0$ is identified on $\boldsymbol{\Theta}$ such that: $G^{(NL)}(\boldsymbol{\beta}) = \mathbf{0}_D$ iff $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ a.s.

Assumption 3.7 (*Asymptotics: Stacked*)**.** N_T is a non-decreasing function of T such that $N_T \to \infty$ as $T \to \infty$ and $T/N_T \to \rho \in [0; \infty)$.

Assumption 3.7 is more restrictive than Assumption 3.4 and ensures that the "many moments" bias is not explosive as $N, T \to \infty$. Such condition is standard in the literature (e.g. Bekker, 1994, Han and Phillips, 2006, and Newey and Windmeijer, 2009).

Let

$$\boldsymbol{\Gamma}_t^{(\xi)} = \lim_{N\to\infty} \mathrm{E}_{\mathcal{F}}\left[\overline{\boldsymbol{\Gamma}}_t^{(\xi)}\right], \tag{43}$$

be the "expected Jacobian" matrix at time t. For example, for the linear approach the above matrix takes the form

$$\boldsymbol{\Gamma}_t^{(L)} = \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d] \left(f_t\, \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t+1}'] - f_{t+1}\, \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t}']\right). \tag{44}$$

The modified influence function associated with the "many moments" setup is given by

$$\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)MMT} = \frac{1}{T_1}\sum_{t=1}^{T_1} \left(\boldsymbol{\Gamma}_t^{(\xi)}\right)' \boldsymbol{\mu}_{i,t}^{(\xi)}. \tag{45}$$

---

19 The same holds true for estimators that make use of $\mathcal{O}(T^2)$ moment conditions, which become naturally available in panels with weakly exogenous regressors. For instance, Lee et al. (2017) show that consistent estimation of the optimal weighting matrix generally requires $T^3/N \to \rho \in [0; \infty)$. Notably, when the objective function involves $\mathcal{O}(T^2)$ moment conditions, the use of a non-optimal weighting matrix may result in an inconsistent GMM estimator (Alvarez and Arellano, 2003).

20 In fact, such practice bears adverse implications for local and global identification, see Satchachai and Schmidt (2008) for more details.

Given the above definition, we assume that the following variance–covariance matrix

$$\boldsymbol{\Omega}_{MMT}^{(\xi)} = \operatorname*{plim}_{T\to\infty} \mathrm{E}_{\mathcal{F}}\left[T_1 \overline{\boldsymbol{\mu}}_{i,T}^{(\xi)MMT}\left(\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)MMT}\right)'\right], \tag{46}$$

is well defined, a.s. finite and $\mathcal{F}$-measurable. Note that because $\mathrm{E}_{\mathcal{D}}[\boldsymbol{\mu}_{i,t}^{(\xi)}] = \mathbf{0}_D$ and all $\boldsymbol{\Gamma}_t^{(\xi)}$ are $\mathcal{F}$-measurable, it follows directly that

$$\mathrm{E}_{\mathcal{D}}\left[\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)MMT}\right] = \frac{1}{T_1}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{D}}\left[\left(\boldsymbol{\Gamma}_t^{(\xi)}\right)'\boldsymbol{\mu}_{i,t}^{(\xi)}\right] = \frac{1}{T_1}\sum_{t=1}^{T_1}\left(\boldsymbol{\Gamma}_t^{(\xi)}\right)'\mathrm{E}_{\mathcal{D}}\left[\boldsymbol{\mu}_{i,t}^{(\xi)}\right] = \mathbf{0}_K. \tag{47}$$

Hence the "many-moments" influence function $\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)MMT}$ inherits the conditional mean and mixing properties from the original version based on "averaging" of the estimating equations.[21] The main result of this section is presented in Theorem 2.

Theorem 2. *Suppose that* $\mathrm{rk}\left[\boldsymbol{\Omega}_{MMT}^{(\xi)}\right] = K$ *a.s. and Assumption 3.5–3.7 hold true, while Assumption 3.1 holds with* $r = 8$. *Then, for* $\xi = \{L; NL\}$:

$$\sqrt{NT_1}\left(\widehat{\boldsymbol{\beta}}_{MMT}^{(\xi)} - \boldsymbol{\beta}_0 - \frac{1}{N}\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)}\right) \Rightarrow \left(\boldsymbol{\Gamma}_{MMT}^{(\xi)}\right)^{-1}\left(\boldsymbol{\Omega}_{MMT}^{(\xi)}\right)^{1/2}\boldsymbol{\psi}_{MMT} \quad (\mathcal{F}-stably), \tag{48}$$

where $\boldsymbol{\Omega}_{MMT}^{(\xi)}$, *defined in Eq. (46), is independent of* $\boldsymbol{\psi}_{MMT} \sim N(\mathbf{0}_K, \boldsymbol{I}_K)$.

Proof. See Appendix A. □

The $\mathcal{F}$-measurable bias term $\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)}$ generally consists of two terms:

$$\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)} = \boldsymbol{b}_{T,\mathcal{F}}^{(\xi)IP} + \boldsymbol{b}_{T,\mathcal{F}}^{(\xi)MMT}, \tag{49}$$

i.e. the "incidental parameters" component, $\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)IP}$, which is zero for the nonlinear approach as well as the linear approach with the "delete-one" correction, and the "many moments" component, $\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)MMT}$, which is non-zero in general. In particular, as it is the case with all standard problems entailing an increasing number of moment conditions, $\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)MMT}$ is determined by the correlation structure between $\boldsymbol{\Gamma}_t^{(\xi)}\boldsymbol{\mu}_{i,t}^{(\xi)}$ and the (individual-specific) influence functions associated with the Jacobian matrix of the estimating equations.

In principle the "many moments" component of the bias can be removed by means of higher order JIVE-type correction (e.g. Angrist et al., 1999). In the present case such correction would require the use of a U-statistic of degree 4, which is akin to network models such as those of Graham (2017) and Jochmans (2017). Alternatively, one can use the two-sample/split-sample approach, as proposed in Angrist and Krueger (1995) and Chernozhukov et al. (2018). In practice, simulation evidence reported in Section 6 suggests that the bias appears to be almost negligible.

4. Extensions

4.1. Additional restrictions: Lack of serial correlation

So far we have restricted our attention to situations where the instrument vector $\boldsymbol{z}_{i,t}$ is given/known. However, under suitable restrictions on the DGP in Eq. (1) an additional set of instruments (moment conditions) can be considered. For example, if $\varepsilon_{i,t}$ is serially uncorrelated, i.e.

$$\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,t}\varepsilon_{i,s}] = 0; \quad \forall t \neq s, \tag{50}$$

an additional set of moment conditions is available for estimation of $\boldsymbol{\beta}$. An assumption of this type is commonly used for fixed T inference in models with weakly exogenous regressors, such as when $\boldsymbol{x}_{i,t}$ contains a lagged dependent variable (see Arellano, 2003). For T large, Eq. (50) can be extended accordingly.

In specific, if condition Eq. (50) is satisfied, then for each time period t the following variable can be used as instrument

$$h_{i,t}(\boldsymbol{\beta}) = y_{i,t+s} - \boldsymbol{x}_{i,t+s}'\boldsymbol{\beta}, \tag{51}$$

for $s = -t+1, \dots, -1, 2, \dots, T-t$. Here we use the notation $h_{i,t}(\boldsymbol{\beta})$ to differentiate between known instruments $\boldsymbol{z}_{i,t}$, and "unknown" ones. In the additive error components structure, the moment conditions in Eq. (51) are usually attributed to Ahn and Schmidt (1995). Ahn et al. (2001) also discuss such moment conditions for a model with a single common factor.

[21] Notice that one of the implications of Assumption 3.1 is $\|\boldsymbol{\Gamma}_t^{(\xi)}\| < \Xi, \forall t$.

Observe that $h_{i,t}(\boldsymbol{\beta})$ is a function of the idiosyncratic and factor components. For example, for $s = -1$,

$$h_{i,t}(\boldsymbol{\beta}_0) = \lambda_i f_{t-1} + \varepsilon_{i,t-1}. \tag{52}$$

Hence, it is clearly seen that both components determine the asymptotic distribution of any estimator that utilizes moment conditions of this type. Identification issues aside, all previous theoretical results accommodate $h_{i,t}(\boldsymbol{\beta})$-type instruments among $\boldsymbol{z}_{i,t}$ from Eq. (51).

4.2. Multiple factors

So far we have assumed for ease of exposition that the number of factors is known and is set equal to $L = 1$. In what follows we consider the generalization of our framework to the case of multiple factors:

$$y_{i,t} = \boldsymbol{\beta}'\boldsymbol{x}_{i,t} + \boldsymbol{\lambda}_i'\boldsymbol{f}_t + \varepsilon_{i,t}, \tag{53}$$

where both $\boldsymbol{\lambda}_i$ and $\boldsymbol{f}_t$ are L-dimensional vectors.

To see how our approach can be extended, consider the following shorthand notation: $\boldsymbol{y}_{i,s:t}$, for $s \le t$, denotes vectors of the form $\boldsymbol{y}_{i,s:t} = \left(y_{i,s}, \dots, y_{i,t}\right)'$. Stacking the $(y_{i,t+1}, \dots, y_{i,t+L})$ observations together leads to

$$\boldsymbol{y}_{i,t+1:t+L}' = \boldsymbol{\beta}'\boldsymbol{X}_{i,t+1:t+L} + \boldsymbol{\lambda}_i'\boldsymbol{F}_{t+1:t+L} + \boldsymbol{\varepsilon}_{i,t+1:t+L}', \tag{54}$$

where $\boldsymbol{F}_{t+1:t+L} = (\boldsymbol{f}_{t+1}, \dots, \boldsymbol{f}_{t+L})$ is $[L \times L]$. Let $|\boldsymbol{F}_{t+1:t+L}|$ denote the determinant of $\boldsymbol{F}_{t+1:t+L}$ and let $\boldsymbol{F}_{t+1:t+L}^{+}$ denote the corresponding adjoint matrix, such that $\boldsymbol{F}_{t+1:t+L}\boldsymbol{F}_{t+1:t+L}^{+} = |\boldsymbol{F}_{t+1:t+L}|$. Then, analogously to the model with $L = 1$, the FQD transformation is given by

$$|\boldsymbol{F}_{t+1:t+L}|(y_{i,t} - \boldsymbol{\beta}'\boldsymbol{x}_{i,t}) - (\boldsymbol{y}_{i,t+1:t+L}' - \boldsymbol{\beta}'\boldsymbol{X}_{i,t+1:t+L})\boldsymbol{F}_{t+1:t+L}^{+}\boldsymbol{f}_t = |\boldsymbol{F}_{t+1:t+L}|\varepsilon_{i,t} - \boldsymbol{\varepsilon}_{i,t+1:t+L}'\boldsymbol{F}_{t+1:t+L}^{+}\boldsymbol{f}_t, \tag{55}$$

for $t = 1, \dots, T - L$, since

$$|\boldsymbol{F}_{t+1:t+L}|\boldsymbol{f}_t - \boldsymbol{F}_{t+1:t+L}\boldsymbol{F}_{t+1:t+L}^{+}\boldsymbol{f}_t = |\boldsymbol{F}_{t+1:t+L}|\boldsymbol{f}_t - |\boldsymbol{F}_{t+1:t+L}|\boldsymbol{f}_t = \boldsymbol{0}_L. \tag{56}$$

Similarly to the single factor case, in general the estimating equations will take the form of a nonlinear *U-statistic* of degree $L+1$, as both $|\boldsymbol{F}_{t+1:t+L}|$ and $\boldsymbol{F}_{t+1:t+L}^{+}\boldsymbol{f}_t$ are products of averages. This holds true unless there exists an L-dimensional vector $\boldsymbol{d}_{i,t}$ such that

$$\boldsymbol{d}_{i,t} = \boldsymbol{\Lambda}_i^d\boldsymbol{f}_t + \boldsymbol{\varepsilon}_{i,t}^d; \quad \mathrm{E}_{\mathcal{F}}[\boldsymbol{\varepsilon}_{i,s}^d | q_{i,t}] = \boldsymbol{0}_L, \tag{57}$$

for $s \ge t$, where $\boldsymbol{\Lambda}_i^d$ is $[L \times L]$.[22] In this case the corresponding estimator remains linear in $\boldsymbol{\beta}$ irrespective of the number of factors.

Note that identification of $\boldsymbol{\beta}$ with multiple factors requires that T is strictly larger than L. This condition, which is only relevant for T fixed, is similar to that in other fixed T GMM or least-squares estimators available in the literature, see e.g. Remark 1 in Juodis and Sarafidis (2020) and Assumptions T-C in Westerlund et al. (2019).

Let L_0 denote the true number of factors. The value of L_0 can be determined using a BIC information criterion, as in Ahn et al. (2013) and Robertson et al. (2018). In particular, define

$$BIC^{(\xi)}(L) = NT_1 \ Q_{NT}(\hat{\boldsymbol{\beta}}_{MM}^{(\xi)}(L)) - \ln(NT_1) \ b(L), \tag{58}$$

where $Q_{NT}(\hat{\boldsymbol{\beta}}_{MM}^{(\xi)}(L))$ is the value of the GMM objective function based on averaged estimating equations and evaluated at $\hat{\boldsymbol{\beta}}_{MM}^{(\xi)}(L)$, $\hat{\boldsymbol{\beta}}_{MM}^{(\xi)}(L)$ denotes the estimate of $\boldsymbol{\beta}$ using L factors, and $b(L)$ is a penalty function that equals a constant times the degrees of freedom of the model. Note that $b(L)$ is strictly decreasing in L.[23] Let

$$\hat{L} = \underset{L=0,\dots,L_{\max}}{\arg\min} \ BIC^{(\xi)}(L), \tag{59}$$

where $L_{\max}$ is such that $L_0 \le L_{max}$. $\hat{L}$ is consistent, i.e. as $N, T \to \infty$, $\hat{L} \xrightarrow{p} L_0$. To see this, let L^- and L^+ denote any two values for L such that $L^- < L_0$ and $L^+ > L_0$. Dropping the superscript (ξ), we have

$$\begin{aligned}
&\mathrm{P}_{\mathcal{D}}\left[BIC(L_0) - BIC(L^+) > 0\right] \\
&= \mathrm{P}_{\mathcal{D}}\left[NT_1 \ \left(Q_{NT}(\hat{\boldsymbol{\beta}}(L_0)) - Q_{NT}(\hat{\boldsymbol{\beta}}(L^+))\right) + \ln(NT_1)\left(b(L^+) - b(L_0)\right) > 0\right] \\
&\le \mathrm{P}_{\mathcal{D}}\left[NT_1\left(Q_{NT}\hat{\boldsymbol{\beta}}(L_0)\right) + \ln(NT_1)\left(b(L^+) - b(L_0)\right) > 0\right] \to 0,
\end{aligned} \tag{60}$$

[22] Alternatively, one can combine a single variable $d_{i,t}$ with L weights, $\boldsymbol{q}_{i,t}$, see Juodis and Sarafidis (2020) for more details.

[23] Alternatively, the second term on the right-hand side of Eq. (58) can be set equal to $h(NT_1)b(L)$, where $h(NT_1) = \ln(\ln(NT_1))$. Such choice is valid given that $h(NT_1)$ satisfies $h(NT_1) \to \infty$ and $h(NT_1)/(NT_1) \to 0$, as $N, T_1 \to \infty$. See Geweke and Meese (1981) and Bai and Ng (2002) for more details.

since the first term is chi-squared distributed by Theorem 1, and thus it is $\mathcal{O}_P(1)$, whereas the second term diverges to $-\infty$. On the other hand,

$$\begin{aligned} &\mathrm{P}_{\mathcal{D}}\left[BIC(L_0) - BIC(L^-) > 0\right] \\ &= \mathrm{P}_{\mathcal{D}}\left[\left(Q_{NT}(\hat{\boldsymbol{\beta}}(L_0)) - Q_{NT}(\hat{\boldsymbol{\beta}}(L^-))\right) + \frac{\ln(NT_1)}{NT_1}\left(b(L^-) - b(L_0)\right) > 0\right] \to 0, \end{aligned} \tag{61}$$

because the first term converges to a fixed negative number (since $Q_{NT}(\hat{\boldsymbol{\beta}}(L_0)) \to 0$ and $Q_{NT}(\hat{\boldsymbol{\beta}}(L^-)) \to c$, $0 < c < \infty$), whereas the second term converges to zero. The same BIC expression is valid for T fixed as well, because the first term in (60) remains bounded, whereas the first term in (61) behaves as before.

Note that a similar BIC method can also be used for the GMM estimator based on stacked estimating equations. In particular, as it is already pointed out by Ahn et al. (2013), the overidentifying restrictions test statistic computed with a non-optimal weighting matrix is a weighted average of independent chi-squared variables, and thereby it remains bounded.[24]

5. Inference and identification

5.1. Variance matrix estimation

Asymptotically valid inference requires consistent estimation of $\boldsymbol{\Gamma}^{(\xi)}$ and $\boldsymbol{\Omega}^{(\xi)}$ for the averaged equations, and $\boldsymbol{\Gamma}^{(\xi)}_{MMT}$ and $\boldsymbol{\Omega}^{(\xi)}_{MMT}$ for the stacked equations. As with any standard GMM problem, $\boldsymbol{\Gamma}^{(\xi)}$ and $\boldsymbol{\Gamma}^{(\xi)}_{MMT}$ can be estimated based on the corresponding sample analogues evaluated at any consistent estimator of $\boldsymbol{\beta}_0$. In particular, for either $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{MM}$ or $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{MMT}$, we suggest

$$\widehat{\boldsymbol{\Gamma}}^{(\xi)}(\widehat{\boldsymbol{\beta}}) = \left[\partial\overline{\widetilde{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'\right]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \tag{62}$$

and

$$\widehat{\boldsymbol{\Gamma}}^{(\xi)}_{MMT}(\widehat{\boldsymbol{\beta}}) = \frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{\Gamma}}^{(\xi)}_t(\widehat{\boldsymbol{\beta}})\right)'\overline{\boldsymbol{\Gamma}}^{(\xi)}_t(\widehat{\boldsymbol{\beta}}); \quad \overline{\boldsymbol{\Gamma}}^{(\xi)}_t(\widehat{\boldsymbol{\beta}}) = \left[\partial\overline{\boldsymbol{m}}^{(\xi)}_t(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'\right]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}. \tag{63}$$

Next, we discuss consistent estimation of $\boldsymbol{\Omega}^{(\xi)}$ and $\boldsymbol{\Omega}^{(\xi)}_{MMT}$. As a building block for this, we consider estimation of the individual components of $\boldsymbol{\mu}^{(\xi)}_{i,t}$, as given by Eq. (19), i.e.

$$\boldsymbol{\mu}^{(\xi)}_{i,t} = \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda^{(\xi)}_i]\boldsymbol{z}_{i,t}\left(f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}\right) - \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]q_{i,t}\left(f_{t+1}\varepsilon^{(\xi)}_{i,t} - f_t\varepsilon^{(\xi)}_{i,t+1}\right), \tag{64}$$

for $\xi \in \{L, NL\}$. Set

$$\widehat{\varepsilon}_{i,t} = y_{i,t} - \boldsymbol{x}'_{i,t}\widehat{\boldsymbol{\beta}}; \quad \widehat{\varepsilon}^{(NL)}_{i,t} = \widehat{\varepsilon}_{i,t}; \quad \widehat{\varepsilon}^{(L)}_{i,t} = d_{i,t}, \tag{65}$$

where $\widehat{\boldsymbol{\beta}}$ is either $\widehat{\boldsymbol{\beta}}_{MM}$ or $\widehat{\boldsymbol{\beta}}_{MMT}$. While $\widehat{\varepsilon}_{i,t}$ is not a suitable estimate for $\varepsilon_{i,t}$ directly, it remains a valid plug-in estimate in the term $f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}$. That is, assuming both f_t and f_{t+1} are known, it is straightforward to see that $f_{t+1}\widehat{\varepsilon}_{i,t} - f_t\widehat{\varepsilon}_{i,t+1}$ is a suitable estimate for $f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}$. The same result holds for the expression involving $\varepsilon^d_{i,t}$; that is, $\widehat{\varepsilon}^d_{i,t}$ is a valid plug-in estimate in $f_{t+1}\varepsilon^d_{i,t} - f_t\varepsilon^d_{i,t+1}$.

In practice, neither f_t or f_{t+1}, nor $\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda^{(\xi)}_i]$ and $E_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]$ are observed. However, separate estimation of these components is not necessary as all expressions involved in the estimation of $\boldsymbol{\Omega}^{(\xi)}$ and $\boldsymbol{\Omega}^{(\xi)}_{MMT}$ are multiplicative in these components.

Define $g^{(\xi)}_{(q),t,s} = \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda^{(\xi)}_i]f_s$ and $\boldsymbol{g}_{(\boldsymbol{z}),t,s} = \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]f_s$ for $t = 1, \ldots, T_1$ and $s = \{t; t+1\}$. The plug-in estimates of these quantities can be constructed as follows:

$$\widehat{g}^{(\xi)}_{(q),t,s} = \frac{1}{N}\sum_{j=1}^{N} q_{j,t}d^{(\xi)}_{j,s}; \quad \widehat{\boldsymbol{g}}_{(\boldsymbol{z}),t,s} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{z}_{j,t}\widehat{\varepsilon}_{j,s}. \tag{66}$$

Intuitively, this is justified because for any $t = 1, \ldots, T$, we have $\widehat{g}^{(\xi)}_{(q),t,s} = g^{(\xi)}_{(q),t,s} + \mathcal{O}_P(N^{-1/2})$. That is, $\widehat{g}^{(\xi)}_{(q),t,s}$ is a consistent estimator for $g^{(\xi)}_{(q),t,s}$ as $N \to \infty$. The same result holds for $\widehat{\boldsymbol{g}}_{(z),t,s}$.

24 However, certain restrictions on the convergence rate of $N, T \to \infty$ are required. See Theorem 3 for more details.

Hence, the plug-in analogue of Eq. (64) is of the form:

$$\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi)} = \boldsymbol{z}_{i,t}\left(\widehat{g}_{(q),t,t+1}^{(\xi)}\widehat{\varepsilon}_{i,t} - \widehat{g}_{(q),t,t}^{(\xi)}\widehat{\varepsilon}_{i,t+1}\right) - q_{i,t}\left(\widehat{\boldsymbol{g}}_{(z),t,t+1}\widehat{\varepsilon}_{i,t}^{(\xi)} - \widehat{\boldsymbol{g}}_{(z),t,t}\widehat{\varepsilon}_{i,t+1}^{(\xi)}\right).$$

This plug-in expression can be used directly to consistently estimate $\boldsymbol{\Omega}^{(\xi)}$. We consider the following (centered) estimator of $\boldsymbol{\Omega}^{(\xi)}$[25]:

$$\widehat{\boldsymbol{\Omega}}^{(\xi)}(\widehat{\boldsymbol{\beta}}) = \frac{1}{NT_1}\sum_{i=1}^{N}\left(\left(\sum_{t=1}^{T}\left(\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi)} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi)}\right)\right)\left(\sum_{t=1}^{T}\left(\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi)} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi)}\right)\right)'\right). \quad (67)$$

For the approach with stacked estimating equations, let

$$\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi),MMT} = \left(\overline{\boldsymbol{\Gamma}}_{t}^{(\xi)}(\widehat{\boldsymbol{\beta}})\right)'\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi)}. \quad (68)$$

The corresponding plug-in estimator is given by

$$\widehat{\boldsymbol{\Omega}}_{MMT}^{(\xi)}(\widehat{\boldsymbol{\beta}}) = \frac{1}{NT_1}\sum_{i=1}^{N}\left(\left(\sum_{t=1}^{T}\left(\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi),MMT} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi),MMT}\right)\right)\left(\sum_{t=1}^{T}\left(\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi),MMT} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{\mu}}_{i,t}^{(\xi),MMT}\right)\right)'\right). \quad (69)$$

When T is fixed, it is straightforward to show consistency of the proposed plug-in estimators $\widehat{\boldsymbol{\Omega}}^{(\xi)}(\widehat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\Omega}}_{MMT}^{(\xi)}(\widehat{\boldsymbol{\beta}})$. On the other hand, a new proof is required for the case where T is large. This is the subject of the following theorem:

Theorem 3. *Let $\widehat{\boldsymbol{\beta}}$ be such that $\sqrt{NT}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \mathcal{O}_P(1)$. Then under the same set of assumptions used in Theorem 1 with $r = 8$, we have*

$$\widehat{\boldsymbol{\Omega}}^{(L)}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\Omega}^{(L)} + o_P(1), \quad \text{given } T/N^2 \to 0. \quad (70)$$

$$\widehat{\boldsymbol{\Omega}}^{(NL)}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\Omega}^{(NL)} + o_P(1). \quad (71)$$

Proof. See Appendix A. □

We note that the restriction $T/N^2 \to 0$ is required only if $\widehat{\boldsymbol{g}}$ is obtained without a "delete-one" correction, where $\widehat{\boldsymbol{g}}$ denotes the vector that collects all $\widehat{\boldsymbol{g}}_{(z),t,t}$, $\widehat{\boldsymbol{g}}_{(z),t,t+1}$, $\widehat{g}_{(q),t,t}^{(\xi)}$ and $\widehat{g}_{(q),t,t+1}^{(\xi)}$ terms, for $t = 1, \dots, T_1$, with the corresponding true parameters denoted by $\boldsymbol{g}_0$; otherwise, this restriction can be dropped. The nonlinear estimator does not require such restriction even in the absence of a "delete-one" correction for $\widehat{\boldsymbol{g}}$. The consistency result for $\widehat{\boldsymbol{\Omega}}_{MMT}^{(\xi)}(\widehat{\boldsymbol{\beta}})$ is similar to that in Theorem 3 albeit at the expense of many additional, asymptotically negligible remainder terms. To save space, we refrain from providing further details.

As an alternative to the asymptotic approximation, one can use the cross-sectional bootstrap as in Kapetanios (2008) or Galvao and Kato (2014), as the estimating equations are asymptotically linear in $\boldsymbol{\mu}_{i,t}^{(\xi)}$. However, while it is reasonable to expect that this bootstrap approach works in our setup, we do not attempt to prove formally the asymptotic validity of it.

5.2. Identification

As it is usually the case with GMM approaches in general, consistency and asymptotic normality of the proposed estimator requires that $\boldsymbol{\beta}$ is locally and globally identified from a given set of moment conditions. We start this section by demonstrating the possibility that identification could fail when the model is estimated based on averaged estimating equations but it could still be achieved when the model is estimated based on stacked estimating equations; the reverse is not true. In other words, when it comes to identification of $\boldsymbol{\beta}_0$, one cannot do worse by using stacked moments as opposed to averaged ones. This is summarized in the following proposition:

Proposition 1. *Under the assumptions employed in Theorem 2, as well as strict stationarity of the underlying time-series mixing process, the class of globally and locally identified models based on averaged estimating equations is no larger than that based on stacked estimating equations.*

Proof. See Appendix A. □

In what follows we shall discuss local identification first, followed by global identification for the nonlinear approach.

[25] The centered estimator of $\boldsymbol{\Omega}^{(\xi)}$ is considered mostly to improve power properties of all test statistics under alternative hypotheses, especially in combination with identification robust inference procedures discussed in Section S.3 of the Supplementary Appendix.

5.2.1. Local identification

Local identification crucially depends on the properties of either $\boldsymbol{\Gamma}^{(\xi)}$ (the limiting Jacobian matrix) for the averaged estimating equations, or matrix $\boldsymbol{\Gamma}^{(\xi)}_{MMT}$ for the stacked estimating equations. In particular, if the full rank condition in Assumptions 3.2 and 3.5 fails, the results provided in Section 3 are invalidated, at least partially. In what follows, we analyze stylized special cases and discuss necessary and sufficient conditions to ensure local identification.

Linear Approach

We consider the original model in Eq. (1) with $K = 1$:

$$y_{i,t} = \beta x_{i,t} + \lambda_i f_t + \varepsilon_{i,t}. \tag{72}$$

It is straightforward to show that the "expected Jacobian matrix" at time t is given by

$$\boldsymbol{\Gamma}^{(L)}_t = \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d]\left(f_t\,\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}x_{i,t+1}] - f_{t+1}\,\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}x_{i,t}]\right). \tag{73}$$

The expression above is non-zero when $\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d] \neq 0$ (relevant weight) and either $\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}x_{i,t+1}] \neq \mathbf{0}_D$ or $\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}x_{i,t}] \neq \mathbf{0}_D$ (relevant instruments). When $x_{i,t}$ is also driven by a common shock, the former condition on its own is sufficient to ensure $\boldsymbol{\Gamma}^{(L)}_t \neq 0$. To see this, let $x_{i,t} = \pi_i f_t^x + \varepsilon_{i,t}^x$, where π_i is independent of $\varepsilon_{i,t}^x$. Setting $\boldsymbol{z}_{i,t} = x_{i,t}$ and $q_{i,t} = x_{i,t-1}$, which implies that $x_{i,t}$ is treated as weakly exogenous, we obtain

$$\begin{aligned}\boldsymbol{\Gamma}^{(L)}_t =& \mathrm{E}_{\mathcal{F}}[\lambda_i^d\pi_i]\,\mathrm{E}_{\mathcal{F}}[\pi_i^2]\left(f_{t-1}^x f_t f_t^x f_{t+1}^x - f_{t-1}^x f_{t+1} g_t^2\right)\\ &+ \mathrm{E}_{\mathcal{F}}[\lambda_i^d\pi_i]\left(f_{t-1}^x f_t\,\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,t}^x\varepsilon_{i,t+1}^x] - f_{t-1}^x f_{t+1}\,\mathrm{E}_{\mathcal{F}}[(\varepsilon_{i,t}^x)^2]\right).\end{aligned} \tag{74}$$

In the case where $f_t^x = f_t$, the first term in the above expression equals zero. However, the second term remains non-zero a.s., so long as $\{f_t\}_{t=1}^T$ are stochastic with a continuous distribution (or non-zero constants).

Remark 9 (*Comparison with the Fixed T GMM Literature*)**.** Similar conditions are required to ensure local identification of β using alternative (fixed T) GMM approaches that treat the factors as parameters. For instance, one can show that the Jacobian matrix at period t for the QLD GMM estimator of Ahn et al. (2013) involves terms of the form $f_s^x f_T$, $s \leq t$, which remain non-zero so long as $\{f_t^x\}_{t=1}^T$ and $\{f_t\}_{t=1}^T$ are stochastic with a continuous distribution.

Consider identification of this model with $f_t^x = f_t$ and $T \to \infty$. We focus on averaged estimating equations first. The limiting Jacobian matrix $\boldsymbol{\Gamma}^{(L)} = \mathrm{plim}_{N,T\to\infty} T_1^{-1}\sum_{t=1}^{T_1}\boldsymbol{\Gamma}^{(L)}_t$ is given by

$$\boldsymbol{\Gamma}^{(L)} = \mathrm{E}_{\mathcal{F}}[\lambda_i^d\pi_i]\left(\mathrm{E}[f_{t-1}f_t]\gamma_x(1) - \mathrm{E}[f_{t-1}f_{t+1}]\gamma_x(0)\right). \tag{75}$$

where $\gamma_x(k) = \mathrm{E}[\varepsilon_{i,t}^x\varepsilon_{i,t-k}^x]$. Thus, provided that $\mathrm{E}_{\mathcal{F}}[\lambda_i^d\pi_i] \neq 0$, identification requires that the factor is serially correlated (assuming $\mathrm{E}[f_t] = 0$). Such restriction is natural in the present context because otherwise lagged values of predetermined regressors (either in levels or first differences) are already valid instruments, even without differencing away the common factor component of the error term. That is, absence of serial correlation in f_t implies that there exists $x_{i,t}$ such that $\mathrm{E}[x_{i,t}(\lambda_i f_s + \varepsilon_{i,s})] = 0$ for some $t < s$. Thus for instance, in the pure AR(1) model without quasi-differencing, the Anderson–Hsiao IV estimator with instruments based on appropriate lagged values of endogenous regressors, remains consistent for T large. Similarly, Arellano–Bond type moment conditions remain valid without any transformation that removes the factor component. Clearly, this is a trivial case.

Remark 10 (*Inference with Reduced Convergence Rate*)**.** It is important to emphasize that even if $\mathrm{E}[f_{t-1}f_t] = \mathrm{E}[f_{t-1}f_{t+1}] = 0$, $\widehat{\boldsymbol{\beta}}^{(L)}_{MM}$ remains consistent, albeit its rate of convergence falls to $\sqrt{N}$. Specifically, from

$$\sqrt{T}\boldsymbol{\Gamma}^{(L)} = \mathrm{E}_{\mathcal{F}}[\lambda_i^d\pi_i]\left(\frac{1}{\sqrt{T}}\sum_{t=2}^{T_1} f_{t-1}f_t\gamma_x(1) - \frac{1}{\sqrt{T}}\sum_{t=2}^{T_1} f_{t-1}f_{t+1}\gamma_x(0)\right) + \mathcal{O}_P(T^{-1/2}), \tag{76}$$

it is clear that $\sqrt{T}\boldsymbol{\Gamma}^{(L)}$ follows a normal asymptotic distribution as $T \to \infty$, as normality is guaranteed by the mixing restriction of Assumption 3.1. Notably, since $\sqrt{T}\boldsymbol{\Gamma}^{(L)}$ and $\boldsymbol{\psi} \sim N(\mathbf{0}_D, \boldsymbol{I}_D)$ are asymptotically independent, which is implied by Assumption 3.1(d), inference remains valid without knowledge of the convergence rate of the estimator (either $\sqrt{NT}$ or $\sqrt{N}$). Thus from a practical point of view, there is no need to know the convergence rate of the estimator in this case in order to conduct asymptotically valid inferences.

Remark 11 (*Connection with Literature on Weak Identification*)**.** The above identification result differs qualitatively from the weak-identification setup in Staiger and Stock (1997), where failure of local identification may lead to inconsistent parameter estimates. By contrast, in the present setup, the restriction $\mathrm{E}[f_{t-1}f_t] = \mathrm{E}[f_{t-1}f_{t+1}] = 0$ represents a case of semi-weak identification (see e.g. Antoine and Renault, 2009). This is because so long as $\mathrm{E}_{\mathcal{F}}[\lambda_i^d\pi_i] \neq 0$, $\boldsymbol{\Gamma}^{(L)}$ diverges from zero at a slower rate than $\mathcal{O}((NT)^{-1/2})$, which is indeed tantamount to semi-weak identification.

When one restricts attention to *time-invariant* weights, q_i, it turns out that identification requires $\mathrm{E}[f_t] \neq 0$. When this restriction is violated, the estimator becomes $\sqrt{N}$-consistent. As it is the case with time-varying weights, this does not affect inferences because the usual standardized tests remain valid.

Next we consider local identification for the stacked estimating equations within the linear approach. Here Assumption 3.5 is the necessary condition ensuring that $\boldsymbol{\Gamma}_{MMT}^{(L)}$ is of full rank. In the case where $f_t^x = f_t$, we have

$$\boldsymbol{\Gamma}_{MMT}^{(L)} = \left(\mathrm{E}_{\mathcal{F}}[\pi_i \lambda_i^d]\right)^2 \left(\mathrm{E}[f_{t-1}^2 f_t^2]\gamma_x^2(1) + \mathrm{E}[f_{t-1}^2 f_{t+1}^2]\gamma_x^2(0) - 2\,\mathrm{E}[f_{t-1}^2 f_t f_{t+1}]\gamma_x(0)\gamma_x(1)\right) > 0. \tag{77}$$

Provided $\mathrm{E}_{\mathcal{F}}[\lambda_i^d \pi_i] \neq 0$, it is clear that identification relies purely on fourth-order cross-moments of unobserved factors. Therefore, $\widehat{\boldsymbol{\beta}}_{MMT}^{(L)}$ remains $\sqrt{NT}$-consistent even in the absence of serial correlation in f_t. This result is consistent with Proposition 1.

Nonlinear Approach

We consider a model with a single regressor, as in Eq. (72). It is straightforward to show that the "expected Jacobian matrix" at time t is given by

$$\boldsymbol{\Gamma}_t^{(NL)} = \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i]\left(f_t\,\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}x_{i,t+1}] - f_{t+1}\,\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}x_{i,t}]\right) + \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]\left(f_{t+1}\,\mathrm{E}_{\mathcal{F}}[x_{i,t}q_{i,t}] - f_t\,\mathrm{E}_{\mathcal{F}}[x_{i,t+1}q_{i,t}]\right). \tag{78}$$

Clearly, while $\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}x_{i,t+1}] = \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}x_{i,t}] = \boldsymbol{0}_D$ implies lack of identification for the linear approach, this is not necessarily true for the nonlinear one. On the other hand, the condition of relevance of weights ($\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i] \neq 0$) remains crucial. To illustrate this point, let $x_{i,t}$, $\boldsymbol{z}_{i,t}$ and $q_{i,t}$ be as in the linear approach before. Thus, in this case $\boldsymbol{\Gamma}_t^{(NL)}$ reduces to

$$\begin{aligned}\boldsymbol{\Gamma}_t^{(NL)} &= \mathrm{E}_{\mathcal{F}}[\lambda_i\pi_i]\left(f_{t-1}f_t\,\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,t}^x\varepsilon_{i,t+1}^x] + f_t f_{t+1}\,\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,t}^x\varepsilon_{i,t-1}^x]\right)\\ &\quad - \mathrm{E}_{\mathcal{F}}[\lambda_i\pi_i]\left(f_{t-1}f_{t+1}\,\mathrm{E}_{\mathcal{F}}\left[\left(\varepsilon_{i,t}^x\right)^2\right] + f_t^2\,\mathrm{E}_{\mathcal{F}}[\varepsilon_{i,t-1}^x\varepsilon_{i,t+1}^x]\right).\end{aligned} \tag{79}$$

Therefore, for $\boldsymbol{\Gamma}_t^{(NL)} \neq 0$ a.s., it is required that $\mathrm{E}_{\mathcal{F}}[\pi_i\lambda_i] \neq 0$. In practice, this is a mild restriction because otherwise $x_{i,t}$ becomes strictly exogenous with respect to the factor component and so the standard OLS estimator remains consistent for T fixed. Notice also that this restriction mirrors the high-level Assumption BA.4-1 in Ahn et al. (2013), which states that the regressors are correlated with the factor loadings in the error term. Thus, for $x_{i,t} = \pi_i f_t^x + \varepsilon_{i,t}^x$, their Assumption BA.4-1 implies $\mathrm{E}_{\mathcal{F}}[\lambda_i\pi_i] \neq 0$.

For T large, local identification based on averaged estimating equations requires serial correlation in f_t, while this is not the case for stacked moment conditions. Such result is identical to that already discussed for the linear approach. Therefore, we refrain from providing any details.

5.2.2. Global identification

The majority of the panel data literature with common factors takes it for granted that the moment conditions globally identify the parameters of interest. This is despite the fact that when the parameter space $\boldsymbol{\Theta}$ is sufficiently large, global identification might fail for nonlinear moment conditions, which are typically employed by existing GMM procedures.[26]

As before, we focus on the single regressor case and study asymptotic properties of

$$\boldsymbol{m}_t^{(NL)}(\beta) = \lim_{N\to\infty}\mathrm{E}_{\mathcal{F}}[\overline{\boldsymbol{m}}_t^{(NL)}(\beta)]. \tag{80}$$

Since the moment conditions are at most quadratic, we obtain

$$\boldsymbol{m}_t^{(NL)}(\beta) = (\beta - \beta_0)\left(\boldsymbol{\Gamma}_t^{(NL)} + \frac{1}{2}\boldsymbol{H}_t^{(NL)}(\beta - \beta_0)\right). \tag{81}$$

From the above expression it is clear the $\beta = \beta_0$ is always a solution, as it should be, given that the model is not misspecified. Furthermore, a second solution exists if and only if $\boldsymbol{\Gamma}_t^{(NL)}$ and $\boldsymbol{H}_t^{(NL)}$ are linearly dependent. We investigate this possibility below using a similar example as in the previous section.

Let $x_{i,t} = \pi_i f_t^x + \varepsilon_{i,t}^x$. To study over-identification we set $\boldsymbol{z}_{i,t} = (x_{i,t}, x_{i,t-1})'$ and $q_{i,t} = x_{i,t-2}$. In this case, the expected Jacobian matrix at time t becomes

$$\boldsymbol{\Gamma}_t^{(NL)} = \mathrm{E}_{\mathcal{F}}[\pi_i\lambda_i]\boldsymbol{A}_{1,t}, \tag{82}$$

while

$$\boldsymbol{H}_t^{(NL)} = -2\left(\mathrm{E}_{\mathcal{F}}[\pi_i^2]\boldsymbol{A}_{2,t} + \boldsymbol{A}_{3,t}\right), \tag{83}$$

where $\boldsymbol{A}_{1,t}$, $\boldsymbol{A}_{2,t}$ and $\boldsymbol{A}_{3,t}$ are defined in Appendix A.1. As it turns out in this specific DGP, $\boldsymbol{\Gamma}_t^{(NL)}$ and $\boldsymbol{H}_t^{(NL)}$ are linearly independent unless $f_t^x = f_t$ and *at the same time* $\varepsilon_{i,t}^x$ is serially uncorrelated; in the former case $\boldsymbol{A}_{1,t} = \boldsymbol{A}_{2,t}$, whereas in

26 See e.g. Hayakawa (2016). As per usual, this issue is alleviated when $D > K$ (overidentified case) because it becomes less likely that the same pseudo-true value $\boldsymbol{\beta}_* \in \boldsymbol{\Theta}$ satisfies all moment conditions.

the latter case $\boldsymbol{A}_{3,t} = \mathbf{0}$. Both restrictions combined imply that

$$\boldsymbol{\Gamma}_t^{(NL)} = -\frac{1}{2}\frac{\mathrm{E}_{\mathcal{F}}[\pi_i\lambda_i]}{\mathrm{E}_{\mathcal{F}}[\pi_i^2]}\boldsymbol{H}_t^{(NL)}. \tag{84}$$

In this case, global identification requires $\mathrm{E}_{\mathcal{F}}[\pi_i\lambda_i] = 0$; otherwise, the second (pseudo-true) value of β equals[27]

$$\beta_* = \beta_0 + \frac{\mathrm{E}_{\mathcal{F}}[\pi_i\lambda_i]}{\mathrm{E}_{\mathcal{F}}[\pi_i^2]}. \tag{85}$$

Remark 12 (*Global Identification with Stacked Estimating Equations*)**.** Eq. (84) holds for each t. Therefore, global identification fails regardless of whether one makes use of averaged or stacked estimating equations.

Although the above observation might seem to be pessimistic at first glance, we note that it only applies to the case where $f_t^x = f_t$ and there exists lack of serial correlation in $\varepsilon_{i,t}^x$; the latter implies that, conditional on common shocks, $x_{i,t}$ is independent over t, which is unlikely in many empirically relevant scenarios. Moreover, the aforementioned lack of global identification can be easily overcome if multiple weights are employed. For instance, given two weights $q_{i,t}^{(1)}$ and $q_{i,t}^{(2)}$ corresponding to different variables altogether, it is sufficient for global identification that $\mathrm{E}_{\mathcal{F}}[q_{i,t}^{(1)}\lambda_i] \neq \mathrm{E}_{\mathcal{F}}[q_{i,t}^{(2)}\lambda_i]$. Such condition is likely to be satisfied when at least one weight is time-varying.

Remark 13 (*Identification for the Panel AR(1) Model*)**.** Section S.4 of the Supplementary Appendix discusses local and global identification for the AR(1) model. An important outcome is that the use of time-invariant weights alone may not be sufficient for global identification when T is large. For this reason, we advise using at least one time-varying weight for more general models, especially with predetermined regressors.

6. Finite sample evidence

6.1. Setup

We focus on a setup that generalizes the model studied in Section 5 and consider the following DGP:

$$\begin{aligned}
&y_{i,t} = \beta x_{i,t} + u_{i,t}; \quad u_{i,t} = \lambda_i f_t + \varepsilon_{i,t};\\
&x_{i,t} = \alpha x_{i,t-1} + \delta y_{i,t-1} + \pi_i f_t + \varepsilon_{i,t}^x;\\
&d_{i,t} = \lambda_i^d f_t + \varepsilon_{i,t}^d;\\
&f_t = \mu_f + f_t^*; \quad f_t^* = \alpha_f f_{t-1}^* + \sqrt{1-\alpha_f^2}\,\varepsilon_t^f;\\
&\lambda_i = \mu_\lambda + \lambda_i^*; \quad \pi_i = \mu_\pi + \phi\lambda_i^* + \sqrt{1-\phi^2}\pi_i^*; \quad \lambda_i^d = \mu_{\lambda_d} + \phi\lambda_i^* + \sqrt{1-\phi^2}\lambda_i^{*d},
\end{aligned} \tag{86}$$

for $t = -6, \ldots, T$. Following existing literature (e.g. Juodis and Sarafidis, 2018), all dynamic processes are initialized in the recent past, such that

$$y_{i,-7} = \lambda_i f_{-7} + \varepsilon_{i,-7}; \quad x_{i,-7} = \pi_i f_{-7} + \varepsilon_{i,-7}^x; \quad f_{-7} = \mu_f + \varepsilon_{-7}^f.$$

All stochastic quantities are drawn in each replication. The individual-specific time-invariant variables, $\boldsymbol{v}_i = (\lambda_i^*, \pi_i^*, \lambda_i^{*d})'$, are mutually independent standard normal variates. However, the factor loadings are allowed to be correlated, with correlation coefficient given by ϕ. Similarly, all time-varying error components are mutually independent standard normal variates, except for $\varepsilon_{i,t}$, the variance of which is determined by the proportion of the variation of the total error, $u_{i,t}$, that is due to the purely idiosyncratic disturbance, $\varepsilon_{i,t}$. In particular, motivated by Norkutė et al. (2021), we consider

$$\vartheta = \frac{\mathrm{var}\left(\varepsilon_{i,t}\right)}{\mathrm{var}\left(u_{i,t}\right)} = \frac{\sigma_\varepsilon^2}{\sigma_u^2} = \frac{\sigma_\varepsilon^2}{\mu_f\sigma_\lambda^2 + \mu_\lambda\sigma_f^2 + \sigma_f^2\sigma_\lambda^2 + \sigma_\varepsilon^2}. \tag{87}$$

Solving in terms of σ_ε^2 yields

$$\sigma_\varepsilon^2 = \left(\mu_f\sigma_\lambda^2 + \mu_\lambda\sigma_f^2 + \sigma_f^2\sigma_\lambda^2\right)\frac{\vartheta}{1-\vartheta}. \tag{88}$$

Following Norkutė et al. (2021), we set $\vartheta \in \{1/4; 3/4\}$; in the former (latter) case, 25% (75%) of the variation in $u_{i,t}$ is due to the variation in $\varepsilon_{i,t}$. We specify $\mu_\lambda = 0$ so that this parameter does not affect the variance of $\varepsilon_{i,t}$ through the computation of ϑ. Moreover, we specify $\mu_\pi = -1$ and $\mu_{\lambda_d} = 1$, which implies that the rank condition for CCE (as well as for the linear GMM estimators) is satisfied. We fix $\alpha = \alpha_f = 0.5$ and $\delta = 0.4$. ϕ alternates such that $\phi = \{0; 0.5; 1\}$, whereas

[27] This result is qualitatively similar to the global identification failure studied in Juodis (2018), where it was also shown that if the regressor is spanned only by f_t, identification fails for linear pseudo panel data models with common (cohort-specific) shocks.

Table 1
Simulation results, $N = 200$.

Designs				L-GMM				L-GMM-S			NL-GMM				NL-GMM-S			BC-QMLE			BC-CCE		
T	ϑ	μ_f	ϕ	Bias	RMSE	t	J	Bias	RMSE	t	Bias	RMSE	t	J	Bias	RMSE	t	Bias	RMSE	t	Bias	RMSE	t
10	.25	2	0	.024	.278	.048	.034	.010	.247	.033	.017	.302	.067	.031	−.051	.256	.056	−.099	.209	.122	−.108	3.36	.580
20	.25	2	0	.013	.255	.054	.037	.013	.208	.035	.005	.274	.075	.033	−.042	.224	.049	−.059	.136	.090	.048	1.66	.421
50	.25	2	0	.008	.234	.051	.036	.026	.182	.034	.002	.227	.080	.038	−.023	.194	.040	−.031	.102	.070	.195	.787	.196
10	.25	2	1	.047	.301	.061	.072	.192	.505	.059	.047	.322	.064	.085	−.049	.377	.080	−.206	.437	.169	−.184	3.45	.602
20	.25	2	1	.023	.253	.051	.046	.253	.502	.055	.032	.254	.054	.062	−.001	.310	.086	−.111	.217	.110	−.026	1.60	.391
50	.25	2	1	.001	.230	.049	.043	.371	.594	.091	.008	.236	.050	.057	.064	.316	.095	−.065	.151	.077	.193	.775	.195
10	.75	2	0	.051	.935	.043	.026	.031	.724	.039	.122	.757	.055	.036	−.070	.688	.058	−8.02	1.04	.727	−.203	13.5	.773
20	.75	2	0	.037	.803	.044	.032	.000	.599	.038	.066	.777	.061	.033	−.070	.602	.045	−.783	1.59	.352	.136	6.07	.626
50	.75	2	0	.002	.751	.046	.033	.004	.514	.036	.036	.718	.063	.034	−.039	.529	.041	−.295	.437	.200	.870	2.42	.391
10	.75	2	1	.149	.853	.074	.046	.166	.719	.038	.380	1.02	.139	.131	.150	.710	.070	−1.92	11.2	.987	−.545	13.6	.785
20	.75	2	1	.097	.773	.060	.046	.242	.673	.046	.295	.973	.110	.104	.236	.665	.073	−4.74	5.55	.810	.025	6.09	.641
50	.75	2	1	.039	.702	.057	.049	.374	.710	.063	.173	.832	.081	.091	.371	.700	.097	−.734	.995	.359	.923	2.46	.395
10	.25	0	0	.011	.494	.045	.036	.001	.349	.037	.007	.383	.026	.055	−.033	.348	.074	−.008	.176	.068	.037	1.42	.471
20	.25	0	0	−.003	.446	.052	.036	.004	.329	.040	.003	.376	.024	.050	−.050	.327	.073	−.002	.146	.061	.136	.672	.293
50	.25	0	0	.008	.402	.056	.040	.013	.342	.042	.001	.375	.025	.055	−.078	.344	.074	−.002	.131	.056	−.008	.361	.124
10	.25	0	1	.033	.635	.057	.058	.048	.744	.050	.022	.589	.040	.052	−.163	.639	.191	−.011	.213	.085	.390	1.38	.468
20	.25	0	1	.023	.662	.049	.048	.081	.844	.048	.018	.613	.040	.054	−.240	.672	.209	.004	.183	.069	.142	.686	.301
50	.25	0	1	.000	.742	.047	.045	.135	.991	.042	.013	.688	.046	.048	−.321	.729	.221	−.003	.167	.065	−.004	.373	.133
10	.75	0	0	.024	1.27	.053	.029	−.013	.811	.040	.020	.901	.039	.026	−.127	.832	.054	−.140	.762	.126	2.79	7.59	.717
20	.75	0	0	.051	1.29	.047	.031	.002	.804	.046	.012	.889	.034	.026	−.151	.827	.050	−.018	.443	.080	1.12	3.26	.546
50	.75	0	0	.024	1.08	.062	.030	.009	.794	.041	.003	.882	.031	.028	−.186	.839	.054	−.009	.361	.054	.219	1.31	.273
10	.75	0	1	.030	1.23	.054	.045	.049	.972	.044	.049	1.24	.046	.080	−.215	1.02	.053	−.473	1.64	.276	2.78	7.50	.717
20	.75	0	1	.046	1.29	.055	.041	.061	1.03	.038	.048	1.09	.043	.070	−.242	1.03	.052	−.086	.613	.130	1.06	3.27	.533
50	.75	0	1	−.009	1.25	.051	.044	.126	1.17	.038	.040	.989	.045	.071	−.166	1.11	.054	−.012	.483	.077	.269	1.28	.251

Notes. The results are based on 4000 Monte Carlo draws. "Bias" corresponds to the mean bias, multiplied by $\sqrt{NT}$; "RMSE" denotes the Root Mean Squared Error, multiplied by $\sqrt{NT}$; "t" denotes the empirical rejection frequencies of the Wald test-statistic under the null $H_0 : \beta = 1$, whereas "J" corresponds to the empirical rejection frequencies of the overidentifying restrictions test statistic (where applicable). In both cases the nominal level is set at 5%. "L-GMM" and "NL-GMM" make use of averaged moment conditions and are defined in Eq. (33); "L-GMM-S" and "NL-GMM-S" are the counterparts based on stacked moment conditions and they are defined in Eq. (41). ϑ denotes the proportion of the variation of the total error that is due to the purely idiosyncratic disturbance, and is defined in Eq. (87). μ_f denotes the expected value of f_t. In all designs presented, we fix $\mu_\lambda = 0$, $\mu_\pi = -1$, $\mu_{\lambda_d} = 1$, $\alpha = \alpha_f = 0.5$ and $\delta = 0.4$, where α (α_f) denotes the autoregressive parameter in the DGP for $x_{i,t}$ (f_t), while δ denotes the coefficient of the lagged value of $y_{i,t}$ in the DGP for $x_{i,t}$.

$\mu_f = \{0; 2\}$. When $\mu_f = 0$, identification of β is feasible only by using time-varying weights; for $\mu_f = 2$ both time-invariant and time-varying weights can be informative. Finally, we consider $N = \{50; 200; 500\}$ and $T = \{10; 20; 50\}$. The number of replications equals 4000 for each design.

6.2. Comments

We study the performance of all GMM estimators developed in the present paper. In particular, we present results for the two-step linear and nonlinear estimators based on averaged estimating equations, denoted as "L-GMM" and "NL-GMM", which are defined in Eq. (33) with $\xi = L$ and $\xi = NL$ respectively. We also present results for "L-GMM-S" and "NL-GMM-S", which denote the linear and nonlinear GMM counterparts based on stacked estimating equations and are defined in Eq. (41). As a benchmark, we also consider two popular least-squares methods, namely the bias-corrected QMLE estimator (BC-QMLE) of Moon and Weidner (2017) and the bias-corrected CCE estimator (BC-CCE) of Pesaran (2006) and Chudik and Pesaran (2015a). The results are reported in Tables 1–3 in terms of mean bias and RMSE (both multiplied by $\sqrt{NT}$), as well as rejection frequencies of the t-test statistic. For the GMM estimators based on averaged moment conditions, we also report results on the overidentifying restrictions J-test statistic. For both test statistics, nominal size is set equal to 5%.

All GMM estimators are implemented using two instruments in each time period, $\boldsymbol{z}_{i,t} = (x_{i,t}, x_{i,t-1})'$, and three sets of weights $\boldsymbol{q}_{i,t} = (1, x_{i,t-1}, x_{i,t-2})'$. The time-invariant weight is used for both instruments, while the time-varying weights are used in pairs such that for $z_{i,t} = x_{i,t}$ we set $q_{i,t} = x_{i,t-1}$, whereas for $z_{i,t} = x_{i,t-1}$ we set $q_{i,t} = x_{i,t-2}$. Thus the total number of moment conditions employed for the estimators that make use of averaged (stacked) estimating equations equals 4 ($4T$). Starting values for the nonlinear GMM estimators are based on the estimates provided by the linear GMM estimators, as well as the two-step (FIVU) GMM estimator developed by Robertson and Sarafidis (2015).

The CCE estimator approximates the factors using cross-sectional averages of all observables, namely $(y_{i,t}, x_{i,t}, d_{i,t})'$. Thus, given our GDP the rank condition is always satisfied. In order to account for the time-series bias due to weak exogeneity of the regressor, we follow the suggestion by Chudik and Pesaran (2015a) and Juodis et al. (2021) and

Table 2
Simulation results, $N = 50$.

Designs				L-GMM				L-GMM-S			NL-GMM				NL-GMM-S			BC-QMLE			BC-CCE		
T	ϑ	μ_f	ϕ	Bias	RMSE	t	J	Bias	RMSE	t	Bias	RMSE	t	J	Bias	RMSE	t	Bias	RMSE	t	Bias	RMSE	t
10	.25	2	0	.034	.307	.067	.024	.039	.255	.026	.020	.278	.079	.017	−.030	.257	.066	−.052	.188	.091	−.110	2.29	.425
20	.25	2	0	.022	.278	.068	.020	.030	.212	.028	.013	.271	.095	.015	−.031	.220	.067	−.030	.126	.069	−.042	1.20	.260
50	.25	2	0	.014	.279	.073	.022	.048	.191	.029	.018	.236	.087	.017	−.009	.195	.056	−.019	.102	.067	.023	.694	.147
10	.25	2	1	.069	.325	.094	.072	.377	.592	.114	.089	.346	.113	.100	.088	.379	.149	−.146	.424	.117	−.031	2.23	.408
20	.25	2	1	.035	.281	.071	.045	.489	.685	.135	.061	.315	.103	.083	.149	.366	.180	−.060	.193	.066	−.060	1.22	.269
50	.25	2	1	−.003	.261	.068	.046	.694	.902	.168	.027	.282	.084	.059	.267	.385	.265	−.037	.144	.067	.027	.706	.156
10	.75	2	0	.068	.932	.062	.013	.011	.767	.029	.112	.659	.067	.016	−.120	.633	.082	−3.99	5.43	.610	−.159	7.24	.608
20	.75	2	0	.023	.941	.065	.017	.017	.588	.028	.093	.747	.072	.018	−.076	.540	.076	−.672	1.60	.243	.002	3.54	.431
50	.75	2	0	.031	.888	.067	.018	.035	.512	.027	.041	.822	.068	.019	−.022	.486	.063	−.162	.382	.142	.418	1.79	.249
10	.75	2	1	.160	.814	.100	.040	.353	.746	.064	.361	.800	.167	.114	.313	.683	.266	−5.13	5.74	.902	−.251	7.27	.610
20	.75	2	1	.089	.781	.091	.035	.454	.780	.078	.355	.862	.153	.091	.420	.725	.315	−2.29	3.14	.585	−.083	3.56	.436
50	.75	2	1	.063	.754	.083	.038	.690	.965	.102	.309	.934	.131	.082	.662	.914	.382	−.429	.858	.195	.411	1.77	.249
10	.25	0	0	.009	.529	.058	.017	.001	.350	.026	.007	.335	.021	.023	−.105	.346	.080	−.002	.173	.073	.155	1.06	.373
20	.25	0	0	.013	.555	.064	.021	.004	.348	.026	−.001	.380	.014	.020	−.111	.342	.077	−.004	.147	.057	.023	.554	.209
50	.25	0	0	.008	.468	.063	.027	.021	.363	.026	.008	.423	.019	.026	−.119	.358	.086	−.004	.132	.058	−.061	.369	.128
10	.25	0	1	.027	.666	.070	.038	.105	.656	.049	.012	.581	.045	.084	−.203	.523	.259	−.010	.211	.078	.149	1.06	.373
20	.25	0	1	.031	.699	.063	.035	.147	.741	.039	.008	.604	.038	.077	−.223	.572	.316	−.005	.181	.063	.037	.580	.238
50	.25	0	1	.036	.819	.059	.036	.284	.923	.045	.011	.590	.038	.078	−.212	.617	.308	.000	.160	.058	−.054	.395	.148
10	.75	0	0	.018	1.42	.061	.020	−.032	.816	.030	.023	.842	.015	.012	−.287	.826	.066	−.124	.714	.107	1.25	4.22	.540
20	.75	0	0	.000	1.39	.073	.020	−.012	.798	.038	.039	.849	.012	.015	−.259	.808	.068	−.028	.428	.077	.455	1.95	.347
50	.75	0	0	−.011	1.28	.071	.021	.000	.832	.033	.016	.667	.014	.017	−.221	.825	.052	−.009	.367	.062	.094	.984	.153
10	.75	0	1	.051	1.31	.080	.035	.100	.896	.038	.125	1.08	.044	.028	−.169	.890	.084	−.229	1.15	.225	1.23	4.07	.532
20	.75	0	1	.037	1.41	.069	.035	.140	.947	.034	.140	1.33	.041	.037	−.110	.889	.084	−.041	.610	.117	.418	1.92	.331
50	.75	0	1	.043	1.58	.071	.037	.274	1.09	.038	.224	1.48	.035	.036	.053	.965	.102	−.006	.477	.076	.071	.976	.151

Notes. See Table 1.

Table 3
Simulation results, $N = 500$.

Designs				L-GMM				L-GMM-S			NL-GMM				NL-GMM-S			BC-QMLE			BC-CCE		
T	ϑ	μ_f	ϕ	Bias	RMSE	t	J	Bias	RMSE	t	Bias	RMSE	t	J	Bias	RMSE	t	Bias	RMSE	t	Bias	RMSE	t
10	.25	2	0	.026	.279	.048	.044	.009	.256	.038	.021	.279	.065	.064	−.047	.262	.052	−.143	.253	.181	−.039	5.09	.715
20	.25	2	0	.009	.247	.050	.043	.013	.209	.041	.008	.259	.071	.056	−.039	.220	.046	−.088	.155	.133	−.026	2.29	.540
50	.25	2	0	.007	.230	.053	.044	.012	.178	.038	.006	.198	.070	.055	−.040	.197	.043	−.051	.109	.093	.326	.957	.298
10	.25	2	1	.028	.291	.053	.075	.109	.475	.047	.026	.270	.044	.072	−.121	.406	.062	−.324	.551	.289	−.073	5.01	.720
20	.25	2	1	.013	.240	.049	.051	.140	.435	.044	.013	.235	.047	.061	−.099	.355	.065	−.177	.271	.200	−.114	2.31	.546
50	.25	2	1	.005	.220	.041	.050	.208	.435	.062	.006	.210	.041	.058	−.026	.317	.076	−.096	.167	.114	.307	.969	.288
10	.75	2	0	.044	.863	.046	.038	−.027	.708	.039	.519	.856	.058	.067	−.127	.742	.047	−12.8	15.7	.808	−.174	2.91	.853
20	.75	2	0	.017	.768	.044	.036	−.003	.581	.041	.057	.643	.044	.053	−.083	.612	.044	−1.08	1.87	.626	−.163	9.20	.745
50	.75	2	0	−.004	.712	.046	.039	.011	.510	.036	.022	.577	.051	.056	−.043	.535	.046	−.460	.556	.376	1.45	3.38	.548
10	.75	2	1	.149	.876	.066	.059	.116	.711	.038	.335	1.01	.094	.069	.072	.713	.057	−17.3	17.5	.996	−.990	2.93	.852
20	.75	2	1	.049	.739	.054	.045	.138	.624	.035	.232	.783	.063	.061	.113	.622	.054	−7.46	8.43	.924	.285	9.20	.748
50	.75	2	1	.032	.683	.049	.044	.214	.597	.045	.131	.507	.043	.067	.208	.597	.058	−1.12	1.32	.644	1.55	3.41	.556
10	.25	0	0	.009	.468	.049	.059	.011	.362	.042	.004	.372	.063	.062	.003	.362	.075	−.017	.176	.071	.632	1.96	.584
20	.25	0	0	−.001	.450	.051	.054	.006	.337	.043	.000	.356	.061	.059	−.007	.338	.077	−.001	.147	.058	.264	.901	.405
50	.25	0	0	−.003	.405	.052	.054	.003	.341	.049	.000	.341	.060	.056	−.016	.342	.072	.001	.130	.056	.035	.406	.161
10	.25	0	1	.005	.648	.044	.068	.028	.763	.040	.002	.623	.055	.053	−.058	.709	.131	−.022	.225	.092	.678	1.99	.613
20	.25	0	1	.007	.636	.048	.050	.025	.870	.046	.001	.745	.051	.054	−.113	.798	.153	−.012	.179	.065	.215	.892	.401
50	.25	0	1	.008	.711	.048	.049	.112	1.01	.041	.001	.784	.051	.058	−.157	.866	.152	.000	.160	.059	.040	.399	.168
10	.75	0	0	.007	1.20	.046	.044	−.004	.812	.045	.002	.833	.052	.058	−.032	.816	.053	−.190	1.02	.187	4.33	11.4	.823
20	.75	0	0	−.025	1.20	.043	.044	−.019	.799	.047	.000	.847	.050	.058	−.068	.814	.046	−.048	.459	.086	1.66	4.98	.664
50	.75	0	0	.010	1.01	.048	.048	−.004	.782	.049	.000	.855	.050	.063	−.102	.821	.047	.002	.362	.052	.441	1.79	.395
10	.75	0	1	.011	1.27	.061	.051	−.007	.990	.049	.006	1.21	.052	.078	−.149	1.02	.052	−.709	2.14	.332	4.36	11.6	.810
20	.75	0	1	.033	1.21	.050	.049	.083	1.08	.050	.004	1.16	.051	.075	−.127	1.07	.049	−.109	.675	.152	1.76	5.04	.662
50	.75	0	1	−.010	1.22	.047	.049	.076	1.19	.043	.000	1.17	.050	.074	−.218	1.14	.051	−.029	.470	.076	.428	1.77	.399

Notes. See Table 1.

we implement the half-panel jackknife approach of Dhaene and Jochmans (2015). Motivated by the empirical findings of Juodis et al. (2021), we do not attempt to correct for the $\mathcal{O}_p(1/N)$ bias of the CCE estimator. On the other hand, the BC-QMLE estimator is implemented using the analytical bias-correction result given by Corollary 4.5 in Moon and Weidner

(2017).[28] The choice of the bandwidth, B, is based on the Monte Carlo results presented in Tables 1–2 in Moon and Weidner (2017); thus, we set $B = \{3, 4, 5\}$ for $T = \{10, 20, 50\}$ respectively.

6.3. Results

At first we focus on the results corresponding to $N = 200$, reported in Table 1.[29]

- *(Estimation)* The linear GMM estimator based on averaged moment conditions, L-GMM, has negligible bias in all designs. A similar result holds for the remaining GMM estimators, albeit the bias is occasionally slightly larger in magnitude. For instance, for $T = 10$, $\vartheta = 0.75$, $\mu_f = 2$ and $\phi = 1$ the finite-sample bias (multiplied by $\sqrt{NT}$) of NL-GMM equals 0.374, which implies an average estimate of β roughly equal to 1.004. In the majority of designs, the bias of BC-QMLE and BC-CCE is larger in magnitude than that of GMM. For BC-QMLE the bias exacerbates when T is relatively small, $\mu_f = 2$ and $\vartheta = 0.75$, in which case most of the variation in the total error is due to the idiosyncratic error component. This result is not surprising, and it is tantamount to saying that numerical estimation of the factors (and factor loadings) requires a large enough signal coming from the factor component of the error. However, we note that the bias of BC-QMLE diminishes quickly as T grows. On the other hand, for BC-CCE the bias manifests mainly when T is relatively small, $\vartheta = 0.75$ but $\mu_f = 0$. As with BC-QMLE, the bias of the estimator tends to diminish in samples with larger values of T.
  In regards to RMSE, the performance of linear and nonlinear GMM is similar. Typically, for small values of T GMM estimators with stacked moment conditions have smaller dispersion than their counterparts that are based on averaged moment conditions. This implies that unless there are substantial differences in bias, L-GMM-S and NL-GMM-S tend to achieve a smaller RMSE value compared to L-GMM and NL-GMM. However, for $T = 50$ such differences mostly disappear. BC-QMLE performs best in terms of RMSE, unless T is small and/or the bias of the estimator is very large. BC-CCE appears to have a substantially larger dispersion compared to the remaining estimators, unless $T = 50$. Therefore, BC-CCE typically performs less well in terms of RMSE, even in those cases where the bias of the estimator is small.
- *(Inference)* For the linear GMM estimators, the size of the t-test is close to its nominal value in all cases, albeit "L-GMM-S" occasionally exhibits some minor downward size distortions. The nonlinear GMM estimators perform satisfactorily as well, with some occasional upward distortions the reflect the finite-sample bias observed in the corresponding designs. BC-QMLE and BC-CCE exhibit severe size distortions in most designs. However, these distortions become smaller with higher values of T, although for BC-CCE empirical size often exceeds 20% even when $T = 50$.[30] Thus in general, inference appears to be more reliable for GMM estimators.
  In terms of the rejection frequencies of the J-test statistic, which is only applicable to GMM based on averaged moment conditions, the performance is satisfactory with only minor size distortions observed.

The results for $N = 50$ and $N = 500$ are plausible and in accordance with the aforementioned observations for $N = 200$. As expected, the performance of all estimators improves (deteriorates) with larger (smaller) values of N. For $N = 50$, BC-QMLE dominates in terms of RMSE even in cases where T is small, unless it has a very large bias. On the other hand, for $N = 500$ the performance of GMM and BC-QMLE is of similar magnitude. In general, in comparison to least-squares based procedures, the performance of the GMM estimators appears to be stable and satisfactory across different designs.

7. Concluding remarks

This paper puts forward a novel Method of Moments approach for factor-augmented panels, which is consistent for any value of T. Our approach is motivated by the increasing availability of panels in which the value of T is not negligible. Currently, existing fixed T GMM procedures require estimation of $\mathcal{O}(T)$ nuisance parameters in order to control for the unobserved factors. Consequently, theoretical analysis of such methods becomes intractable even for moderate values of T.

The proposed approach gives rise to estimators that are free from incidental parameters by construction. In particular, we combine two key elements: (i) a quasi-differencing transformation that removes the unobserved *factor loadings* from the error; and (ii) an approximation of the unknown *factors*, based either on observed data or on the composite error term of the model. The latter has an exact factor structure once evaluated at the true value of the slope parameters. Essentially, these two elements allow us to estimate explicitly a fixed number of parameters, regardless of the size of N or T.

We put forward two alternative GMM estimators; one is based on a constant number of "averaged moment conditions" à la Anderson and Hsiao (1982), whereas the other one makes use of "stacked moment conditions", the total number of which increases at the rate of $\mathcal{O}(T)$. We demonstrate that the former estimator is consistent and asymptotically

[28] We are grateful to Martin Weidner for providing to us the computational algorithm for the BC-QMLE estimator.

[29] To save space, we do not present results for the intermediate case $\phi = 0.5$ below. These results are available upon request.

[30] For this reason, it is recommended to use bootstrap-based inference for CCE, as suggested in Juodis et al. (2021) in order to fix some of these distortions.

mixed-normal as $N \to \infty$ for any value of T. The latter remains consistent and asymptotically mixed-normal, although, unsurprisingly, it can be subject to an asymptotic bias proportional to T/N due to the use of "many moment conditions".

The proposed approach can be extended to a wide range of models, motivated by either the micro- or macro-econometric literature. These include non-parametric models (e.g. Su and Jin, 2012), models with spatial dependence (Kuersteiner and Prucha, 2020), unit root tests (Robertson et al., 2018), smooth transition and structural breaks (Qian and Su, 2016), inference in partially identified panels with common factors (e.g. Hong et al., 2019), to mention only a few.

Appendix A. Proofs

A.1. Definition of matrices in the main text

The matrices $\boldsymbol{A}_{1,t}$, $\boldsymbol{A}_{2,t}$ and $\boldsymbol{A}_{3,t}$, introduced in Section 5.2.2 of the main text, are defined as follows:

$$\boldsymbol{A}_{1,t} = \begin{bmatrix} f^x_{t-2}f_t \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t}\varepsilon^x_{i,t+1}] + f^x_t f_{t+1} \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t}] - f^x_{t-2}f_{t+1} \operatorname{E}_{\mathcal{F}}[(\varepsilon^x_{i,t})^2] - f^x_t f_t \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t+1}] \\ f^x_{t-2}f_t \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-1}\varepsilon^x_{i,t+1}] + f^x_{t-1}f_{t+1} \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t}] - f^x_{t-2}f_{t+1} \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-1}\varepsilon^x_{i,t}] - f^x_{t-1}f_t \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t+1}] \end{bmatrix}; \tag{A.1}$$

$$\boldsymbol{A}_{2,t} = \begin{bmatrix} f^x_{t-2}f^x_t \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t}\varepsilon^x_{i,t+1}] + f^x_t f^x_{t+1} \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t}] - f^x_{t-2}f^x_{t+1} \operatorname{E}_{\mathcal{F}}[(\varepsilon^x_{i,t})^2] - f^x_t f^x_t \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t+1}] \\ f^x_{t-2}f^x_t \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-1}\varepsilon^x_{i,t+1}] + f^x_{t-1}f^x_{t+1} \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t}] - f^x_{t-2}f^x_{t+1} \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-1}\varepsilon^x_{i,t}] - f^x_{t-1}f^x_t \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t+1}] \end{bmatrix}; \tag{A.2}$$

$$\boldsymbol{A}_{3,t} = \begin{bmatrix} \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t}] \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t}\varepsilon^x_{i,t+1}] - \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t+1}] \operatorname{E}_{\mathcal{F}}[(\varepsilon^x_{i,t})^2] \\ \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t}] \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-1}\varepsilon^x_{i,t+1}] - \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-2}\varepsilon^x_{i,t+1}] \operatorname{E}_{\mathcal{F}}[\varepsilon^x_{i,t-1}\varepsilon^x_{i,t}] \end{bmatrix}. \tag{A.3}$$

$\boldsymbol{A}_{1,t}$ represents the Jacobian matrix of the moment conditions at time t, divided by the scalar $\operatorname{E}_{\mathcal{F}}[\pi_i\lambda_i]$, for the model considered in Section 5.2.2. $\boldsymbol{A}_{2,t}$ and $\boldsymbol{A}_{3,t}$ enter into the Hessian matrix of the moment conditions for the same model.

A.2. Notation

Before proceeding with derivations define:

$$\Delta_f \varepsilon^{(\xi)}_{i,t+1} = f_t \varepsilon^{(\xi)}_{i,t+1} - f_{t+1}\varepsilon^{(\xi)}_{i,t}, \tag{A.4}$$

$\xi \in \{L; NL\}$, where $\varepsilon^{(L)}_{i,s} = \varepsilon^d_{i,s}$, $\varepsilon^{(NL)}_{i,s} = \varepsilon_{i,s}$, for $s = \{t; t+1\}$.

A.3. Auxiliary lemmas

Lemma 2. *Let $\{\psi_{i,t}\}_{i=1,t=1}^{N,T}$ and $\{\zeta_{i,t}\}_{i=1,t=1}^{N,T}$ be two $\mathcal{F}$ conditionally independent random sequences such that: (i) $\operatorname{E}_{\mathcal{D}}[\psi_{i,t}] = 0$, and $\operatorname{E}_{\mathcal{D}}[\zeta_{i,t}] = 0$; (ii) $\operatorname{E}_{\mathcal{F}}[|\psi_{i,t}|^2] < \Xi$ and $\operatorname{E}_{\mathcal{D}}[|\zeta_{i,t}|^{2+\delta}] < \Xi$, $\delta > 0$; (iii) $\{\zeta_{i,t}\}_{i=1,t=1}^{N,T}$ is a $\mathcal{D}$-conditional α-mixing sequence satisfying $\sum_{m=0}^{\infty} \alpha_i(m)^{1-\frac{2}{2+\delta}}$. Then for all (i,j) with $i \neq j$:*

$$\operatorname{E}_{\mathcal{D}}[\psi_{i,t}\zeta_{j,t}] = 0, \tag{A.5}$$

$$\operatorname{E}_{\mathcal{D}}\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\psi_{i,t}\zeta_{j,t}\right)^2\right] < \Xi, \tag{A.6}$$

for some finite constant Ξ.

Lemma 3. *Let $\{\psi_{i,t}\}_{i=1,t=1}^{N,T}$ be $\mathcal{F}$ conditionally independent random sequences such that $\operatorname{E}_{\mathcal{F}}[\psi_{i,t}] = 0$ and $\operatorname{E}_{\mathcal{F}}[|\psi_{i,t}|^r] < \Xi$. Then, as $N, T \to \infty$ with $T/N \to \rho \in [0; \infty)$,*

$$\sqrt{NT}\frac{1}{T}\sum_{t=1}^{T}(\overline{\psi}_t)^r = o_P(1), \tag{A.7}$$

for $r \in \{3; 4\}$, where $\overline{\psi}_t = N^{-1}\sum_{i=1}^{N}\psi_{i,t}$.

Remark 14. Using the generalized Hölder's inequality, this conclusion applies to any collection of random variables $(\psi^{(1)}_{i,t}, \psi^{(2)}_{i,t}, \psi^{(3)}_{i,t}, \psi^{(4)}_{i,t})$.

Proof of Lemma 2. The first claim follows immediately by conditional $\mathcal{F}$ independence and $\mathrm{E}_{\mathcal{D}}[\zeta_{i,t}] = 0$. As for the second claim, observe that:

$$\begin{aligned}
\mathrm{E}_{\mathcal{D}}\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\psi_{i,t}\zeta_{j,t}\right)^2\right] &= \left|\mathrm{E}_{\mathcal{D}}\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\psi_{i,t}\zeta_{j,t}\right)^2\right]\right| \\
&= \left|\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{E}_{\mathcal{D}}[\psi_{i,t}\zeta_{j,t}\psi_{i,s}\zeta_{j,s}]\right| \\
&= \left|\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{E}_{\mathcal{D}}[\psi_{i,t}\psi_{i,s}]\,\mathrm{E}_{\mathcal{D}}[\zeta_{j,t}\zeta_{j,s}]\right| \\
&\le \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\Xi\left|\mathrm{E}_{\mathcal{D}}[\zeta_{j,t}\zeta_{j,s}]\right| \\
&< \Xi C. \qquad (A.8)
\end{aligned}$$

Here in the third line we make use of conditional independence. The fourth line follows from the triangle inequality and application of the Cauchy–Schwarz inequality using the moment restriction on $\mathrm{E}_{\mathcal{D}}[|\psi_{i,t}|^2] < \Xi$. Finally, the last line follows by the Davydov's inequality, where C is a constant that depends on $\alpha(m)$ and $\mathrm{E}_{\mathcal{D}}[|\zeta_{i,t}|^{4+\delta}]$ only. □

Proof of Lemma 3. We prove this result using the Markov's inequality. In particular, it is sufficient to show that

$$\mathrm{E}_{\mathcal{F}}\left[\left|\sqrt{NT}\frac{1}{T}\sum_{t=1}^{T}(\overline{\psi}_t)^r\right|\right] \to 0, \qquad (A.9)$$

as $N, T \to \infty$. To show this we will make use of several known inequalities. In particular, by the triangle inequality:

$$\mathrm{E}_{\mathcal{F}}\left[\left|\sqrt{NT}\frac{1}{T}\sum_{t=1}^{T}(\overline{\psi}_t)^r\right|\right] \le (NT)^{1/2}\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}_{\mathcal{F}}[|\overline{\psi}_t|^r]. \qquad (A.10)$$

Next, observe that because $\mathrm{E}_{\mathcal{F}}[\psi_{i,t}] = 0$ and $\mathrm{E}_{\mathcal{F}}[|\psi_{i,t}|^r] < \Xi$ the Rosenthal's inequality is applicable so that:

$$\mathrm{E}_{\mathcal{F}}[|\overline{\psi}_t|^r] \le \Xi N^{-r/2}. \qquad (A.11)$$

Collecting all terms:

$$\mathrm{E}_{\mathcal{F}}\left[\left|\sqrt{NT}\frac{1}{T}\sum_{t=1}^{T}(\overline{\psi}_t)^r\right|\right] \le \Xi\left(\frac{T}{N}\right)^{\frac{1}{2}} N^{-\frac{(r-2)}{2}}. \qquad (A.12)$$

The conclusion follows immediately after imposing the condition on T/N, as $r > 2$. □

A.4. Averaged moment conditions

Proof of Lemma 1. The estimating equations can be written as follows:

$$\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[w_{j,t,t+1}^{(\xi)}\boldsymbol{z}_{i,t}\left(y_{i,t} - \boldsymbol{\beta}'\boldsymbol{x}_{i,t}\right) - w_{j,t,t}^{(\xi)}\boldsymbol{z}_{i,t}\left(y_{i,t+1} - \boldsymbol{\beta}'\boldsymbol{x}_{i,t+1}\right)\right], \qquad (A.13)$$

where

$$w_{i,t,t}^{(\xi)} = q_{i,t}\left(\lambda_i^{(\xi)}f_t + \varepsilon_{i,t}^{(\xi)}\right); \quad w_{i,t,t+1}^{(\xi)} = q_{i,t}\left(\lambda_i^{(\xi)}f_{t+1} + \varepsilon_{i,t+1}^{(\xi)}\right),$$

with $\lambda_i^{(L)} = \lambda_i^d$, $\lambda_i^{(NL)} = \lambda_i$, $\varepsilon_{i,t}^{(L)} = \varepsilon_{i,t}^d$ and $\varepsilon_{i,t}^{(NL)} = \varepsilon_{i,t}$. Evaluating (A.13) at $\boldsymbol{\beta}_0$, we can expand this expression as

$$\begin{aligned}
\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(w_{j,t,t+1}^{(\xi)}\boldsymbol{z}_{i,t}\left(\varepsilon_{i,t} + \lambda_i f_t\right) - w_{j,t,t}^{(\xi)}\boldsymbol{z}_{i,t}\left(\varepsilon_{i,t+1} + \lambda_i f_{t+1}\right)\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left(w_{i,t,t+1}^{(\xi)}\boldsymbol{z}_{i,t}\left(\varepsilon_{i,t} + \lambda_i f_t\right) - w_{i,t,t}^{(\xi)}\boldsymbol{z}_{i,t}\left(\varepsilon_{i,t+1} + \lambda_i f_{t+1}\right)\right)
\end{aligned}$$

$$+ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(w_{j,t,t+1}^{(\xi)} \boldsymbol{z}_{i,t} \left(\varepsilon_{i,t} + \lambda_i f_t \right) - w_{j,t,t}^{(\xi)} \boldsymbol{z}_{i,t} \left(\varepsilon_{i,t+1} + \lambda_i f_{t+1} \right) \right)$$
$$= \overline{\boldsymbol{m}}_t^{(\xi)(1)} + \overline{\boldsymbol{m}}_t^{(\xi)(2)}. \tag{A.14}$$

Let us consider the expectation of the second component first:

$$\begin{aligned}
\mathrm{E}_{\mathcal{D}}[\overline{\boldsymbol{m}}_t^{(\xi)(2)}] &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(\mathrm{E}_{\mathcal{D}}[w_{j,t,t+1}^{(\xi)}]\, \mathrm{E}_{\mathcal{D}}[\boldsymbol{z}_{i,t} \left(\varepsilon_{i,t} + \lambda_i f_t \right)] - \mathrm{E}_{\mathcal{D}}[w_{j,t,t}^{(\xi)}]\, \mathrm{E}_{\mathcal{D}}[\boldsymbol{z}_{i,t} \left(\varepsilon_{i,t+1} + \lambda_i f_{t+1} \right)] \right) \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(\mathrm{E}_{\mathcal{D}}[q_{j,t} \lambda_j^{(\xi)} f_{t+1}]\, \mathrm{E}_{\mathcal{D}}[\boldsymbol{z}_{i,t} \lambda_i f_t] - \mathrm{E}_{\mathcal{D}}[q_{j,t} \lambda_j^{(\xi)} f_t]\, \mathrm{E}_{\mathcal{D}}[\boldsymbol{z}_{i,t} \lambda_i f_{t+1}] \right) \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} f_t f_{t+1} \left(\mathrm{E}_{\mathcal{D}}[q_{j,t} \lambda_j^{(\xi)}]\, \mathrm{E}_{\mathcal{D}}[\boldsymbol{z}_{i,t} \lambda_i] - \mathrm{E}_{\mathcal{D}}[q_{j,t} \lambda_j^{(\xi)}]\, \mathrm{E}_{\mathcal{D}}[\boldsymbol{z}_{i,t} \lambda_i] \right) \\
&= \boldsymbol{0}_D,
\end{aligned} \tag{A.15}$$

recalling that $D_t = D$. The first equality makes use of the random sampling assumption, whereas for the second equality we use (Appendix A.4), together with Assumption 3.1 (d). The existence (and boundedness) o f $\mathrm{E}_{\mathcal{D}}[\boldsymbol{z}_{i,t}\lambda_i]$ and $\mathrm{E}_{\mathcal{D}}[q_{j,t}\lambda_i^{(\xi)}]$ is guaranteed by Assumption 3.1 (b). Finally, the last two equalities follow by measurability of f_t with respect to $\mathcal{D}$.

Next, we turn our attention to the first component, which contains all "own" terms $i = j$. Using Eq. (Appendix A.4) we expand $\overline{\boldsymbol{m}}_t^{(\xi)(1)}$ as follows:

$$\overline{\boldsymbol{m}}_t^{(\xi)(1)} = \sum_{s=1}^{8} \overline{\boldsymbol{m}}_t^{(\xi)(1.s)}, \tag{A.16}$$

where

$$\overline{\boldsymbol{m}}_t^{(\xi)(1.1)} = \frac{1}{N^2} \sum_{i=1}^{N} q_{i,t} \lambda_i^{(\xi)} f_{t+1} \boldsymbol{z}_{i,t} \varepsilon_{i,t}; \qquad \overline{\boldsymbol{m}}_t^{(\xi)(1.2)} = \frac{1}{N^2} \sum_{i=1}^{N} q_{i,t} \lambda_i^{(\xi)} f_{t+1} \boldsymbol{z}_{i,t} \lambda_i f_t;$$
$$\overline{\boldsymbol{m}}_t^{(\xi)(1.3)} = \frac{1}{N^2} \sum_{i=1}^{N} q_{i,t} \varepsilon_{i,t+1}^{(\xi)} \boldsymbol{z}_{i,t} \varepsilon_{i,t}; \qquad \overline{\boldsymbol{m}}_t^{(\xi)(1.4)} = \frac{1}{N^2} \sum_{i=1}^{N} q_{i,t} \varepsilon_{i,t+1}^{(\xi)} \boldsymbol{z}_{i,t} \lambda_i f_t;$$
$$\overline{\boldsymbol{m}}_t^{(\xi)(1.5)} = -\frac{1}{N^2} \sum_{i=1}^{N} q_{i,t} \lambda_i^{(\xi)} f_t \boldsymbol{z}_{i,t} \varepsilon_{i,t+1}; \qquad \overline{\boldsymbol{m}}_t^{(\xi)(1.6)} = -\frac{1}{N^2} \sum_{i=1}^{N} q_{i,t} \lambda_i^{(\xi)} f_t \boldsymbol{z}_{i,t} \lambda_i f_{t+1};$$
$$\overline{\boldsymbol{m}}_t^{(\xi)(1.7)} = -\frac{1}{N^2} \sum_{i=1}^{N} q_{i,t} \varepsilon_{i,t}^{(\xi)} \boldsymbol{z}_{i,t} \varepsilon_{i,t+1}; \qquad \overline{\boldsymbol{m}}_t^{(\xi)(1.8)} = -\frac{1}{N^2} \sum_{i=1}^{N} q_{i,t} \varepsilon_{i,t}^{(\xi)} \boldsymbol{z}_{i,t} \lambda_i f_{t+1}.$$

It is easily seen that $\overline{\boldsymbol{m}}_t^{(\xi)(1.2)} + \overline{\boldsymbol{m}}_t^{(\xi)(1.6)} = \boldsymbol{0}$. Moreover, for the nonlinear estimator specifically, we also have

$$\overline{\boldsymbol{m}}_t^{(NL)(1.1)} + \overline{\boldsymbol{m}}_t^{(NL)(1.8)} = \boldsymbol{0}_D; \tag{A.17}$$
$$\overline{\boldsymbol{m}}_t^{(NL)(1.3)} + \overline{\boldsymbol{m}}_t^{(NL)(1.7)} = \boldsymbol{0}_D; \tag{A.18}$$
$$\overline{\boldsymbol{m}}_t^{(NL)(1.4)} + \overline{\boldsymbol{m}}_t^{(NL)(1.5)} = \boldsymbol{0}_D. \tag{A.19}$$

Hence, in this case $\overline{\boldsymbol{m}}_t^{(NL)(1)} = \boldsymbol{0}_D$.

On the other hand, for the linear the remaining 6 terms have non-negligible expectations. At the same time one can show that all conditional and unconditional expectations are well defined. Take $\overline{\boldsymbol{m}}_t^{(L)(1.3)}$ as an example. For each row of $\overline{m}_t^{(L)(1.3)(p)}$, $p = 1, \ldots, D$, we have

$$\mathrm{E}_{\mathcal{F}}[|\overline{m}_t^{(L)(1.3)(p)}|] \leq \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{E}_{\mathcal{F}}[|q_{i,t} \varepsilon_{i,t+1}^{d} z_{i,t}^{(p)} \varepsilon_{i,t}|] \leq \frac{1}{N^2} \sum_{i=1}^{N} \Xi = N^{-1} \Xi = \mathcal{O}(N^{-1}), \tag{A.20}$$

where the second inequality follows from the generalized Hölder's inequality and the fact that all elements have a finite $4 + \delta$ moment. Boundedness of conditional expectations also implies that the corresponding unconditional expectations are bounded. Similarly, it can be shown that $\mathrm{E}_{\mathcal{F}}[|\overline{m}_t^{(L)(1.s)(p)}|] = \mathcal{O}(N^{-1})$ for $s = 1, 4, 5, 7, 8$, $p = 1, \ldots, D$. Combining all terms together, we obtain

$$\mathrm{E}[\overline{\boldsymbol{m}}_t^{(L)(1)}] = \mathcal{O}(N^{-1}). \tag{A.21}$$

Notice that some of these expectations are zero if one restricts attention to time-invariant (strictly exogenous) weights. For example:

$$\mathrm{E}_{\mathcal{D}}\left[\overline{\boldsymbol{m}}_t^{(L)(1.1)}\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{E}_{\mathcal{D}}\left[q_i\lambda_i^d f_{t+1}\boldsymbol{z}_{i,t}\,\mathrm{E}_{\mathcal{D}}[\varepsilon_{i,t}|\boldsymbol{z}_{i,t},\boldsymbol{v}_i]\right] = \boldsymbol{0}_D; \tag{A.22}$$

$$\mathrm{E}_{\mathcal{D}}\left[\overline{\boldsymbol{m}}_t^{(L)(1.5)}\right] = -\frac{1}{N^2}\sum_{i=1}^{N}\mathrm{E}_{\mathcal{D}}\left[q_i\lambda_i^d f_t\boldsymbol{z}_{i,t}\,\mathrm{E}_{\mathcal{D}}[\varepsilon_{i,t+1}|\boldsymbol{z}_{i,t},\boldsymbol{v}_i]\right] = \boldsymbol{0}_D. \quad \square \tag{A.23}$$

Proof of Theorem 1. We break down the proof of this theorem into four distinct steps:

1. Establish negligibility of the bias term;
2. Derive the leading term of the asymptotic expansion;
3. Show consistency of the estimator;
4. Derive asymptotic distribution of the estimator.

To avoid notational clutter, we set $D = K = 1$, unless specified otherwise. In what follows we extensively use the mixing inequalities due to Davydov and Yokoyama, see e.g. Section 1.4 in Doukhan (1994).

Step 1. Using the decomposition in Lemma 1 we express:

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}_0) = \widetilde{\overline{\boldsymbol{m}}}^{(\xi)(1)} + \widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2)}. \tag{A.24}$$

Due to the fact that we use the jackknifed version of the objective function, the first component $\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(1)}$ containing own-terms $i = j$ is zero by construction.

Step 2. Expand

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2)} = \sum_{s=1}^{8}\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.s)}, \tag{A.25}$$

where

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.1)} = \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\lambda_j^{(\xi)} f_{t+1}\boldsymbol{z}_{i,t}\varepsilon_{i,t}; \qquad \widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.2)} = \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\lambda_j^{(\xi)} f_{t+1}\boldsymbol{z}_{i,t}\lambda_i f_t;$$

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.3)} = \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\varepsilon_{j,t+1}^{(\xi)}\boldsymbol{z}_{i,t}\varepsilon_{i,t}; \qquad \widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.4)} = \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\varepsilon_{j,t+1}^{(\xi)}\boldsymbol{z}_{i,t}\lambda_i f_t;$$

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.5)} = -\frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\lambda_j^{(\xi)} f_t\boldsymbol{z}_{i,t}\varepsilon_{i,t+1}; \qquad \widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.6)} = -\frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\lambda_j^{(\xi)} f_t\boldsymbol{z}_{i,t}\lambda_i f_{t+1};$$

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.7)} = -\frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\varepsilon_{j,t}^{(\xi)}\boldsymbol{z}_{i,t}\varepsilon_{i,t+1}; \qquad \widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.8)} = -\frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\varepsilon_{j,t}^{(\xi)}\boldsymbol{z}_{i,t}\lambda_i f_{t+1}.$$

Clearly, $\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.2)} + \widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.6)} = 0$. We combine the remaining 6 terms into three distinct pairs:

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.1+2.5)} = -\frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\lambda_j^{(\xi)}\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1}; \tag{A.26}$$

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.3+2.7)} = \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} \boldsymbol{z}_{i,t}\left(q_{j,t}\varepsilon_{j,t+1}^{(\xi)}\varepsilon_{i,t} - q_{j,t}\varepsilon_{j,t}^{d}\varepsilon_{i,t+1}\right); \tag{A.27}$$

$$\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.4+2.8)} = \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} \lambda_i\boldsymbol{z}_{i,t}q_{j,t}\Delta_f\varepsilon_{j,t+1}^{(\xi)} = \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} \lambda_j\boldsymbol{z}_{j,t}q_{i,t}\Delta_f\varepsilon_{i,t+1}^{(\xi)}. \tag{A.28}$$

Firstly, we have

$$\begin{aligned}\widetilde{\overline{\boldsymbol{m}}}^{(\xi)(2.1+2.5)} &= -\frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1} q_{j,t}\lambda_j^{(\xi)}\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1}\\ &= -\frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{F}}[q_{j,t}\lambda_j^{(\xi)}]\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1} - \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1}\Delta g_{(q),j,t}^{(\xi)}\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1}\end{aligned}$$

$$= -\frac{1}{NT_1}\sum_{i=1}^{N}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^{(\xi)}]\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1} - \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1}\Delta g_{(q),j,t}^{(\xi)}\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1}$$
$$= \widetilde{\widetilde{\boldsymbol{m}}}_{(1)}^{(\xi)(2.1+2.5)} + \widetilde{\widetilde{\boldsymbol{m}}}_{(2)}^{(\xi)(2.1+2.5)}. \tag{A.29}$$

Here we define $\Delta g_{(q),j,t}^{(\xi)} = q_{j,t}\lambda_j^{(\xi)} - \mathrm{E}_{\mathcal{F}}[q_{j,t}\lambda_j^{(\xi)}]$. We show that once scaled by $\sqrt{NT_1}$ the second component is asymptotically negligible, i.e.

$$\sqrt{NT_1}\ \widetilde{\widetilde{\boldsymbol{m}}}_{(2)}^{(\xi)(2.1+2.5)} = o_P(1). \tag{A.30}$$

To show this we will make use of Lemma 2. In particular, as $\mathrm{E}_{\mathcal{F}}[\Delta g_{(q),j,t}^{(\xi)}] = 0$ and

$$\mathrm{E}_{\mathcal{D}}\left[\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1}\right] = \mathrm{E}_{\mathcal{D}}\left[\boldsymbol{z}_{i,t}\,\mathrm{E}_{\mathcal{D}}[\Delta_f\varepsilon_{i,t+1}|\boldsymbol{v}_i,\boldsymbol{z}_{i,t}]\right] = 0, \tag{A.31}$$

the first assertion of that lemma holds. To make use of the second result we note that $\widetilde{\widetilde{\boldsymbol{m}}}_{(2)}^{(\xi)(2.1+2.5)}$ can be symmetrized as follows:

$$\widetilde{\widetilde{\boldsymbol{m}}}_{(2)}^{(\xi)(2.1+2.5)} = \frac{1}{NN_1\sqrt{T_1}}\sum_{i=2}^{N}(\boldsymbol{q}_{i,N}^{(\xi)(1)} + \boldsymbol{q}_{i,N}^{(\xi)(2)}) = \frac{1}{NN_1\sqrt{T_1}}\sum_{i=2}^{N}\sum_{j<i}(\boldsymbol{q}_{i,j}^{(\xi)(1)} + \boldsymbol{q}_{i,j}^{(\xi)(2)}) = \frac{1}{NN_1\sqrt{T_1}}\sum_{i=2}^{N}\sum_{j<i}\boldsymbol{q}_{i,j}^{(\xi)}. \tag{A.32}$$

Here we define

$$\boldsymbol{q}_{i,j}^{(\xi)(1)} = \frac{1}{T_1}\sum_{t=1}^{T_1}\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1}\Delta g_{(q),j,t}^{(\xi)}; \tag{A.33}$$

$$\boldsymbol{q}_{i,j}^{(\xi)(2)} = \frac{1}{T_1}\sum_{t=1}^{T_1}\boldsymbol{z}_{j,t}\Delta_f\varepsilon_{j,t+1}\Delta g_{(q),i,t}^{(\xi)}. \tag{A.34}$$

Moreover, irrespective of the value for T it follows that $\mathrm{E}_{\mathcal{D}}[\boldsymbol{q}_{i,N}^{(\xi)(1)}\boldsymbol{q}_{j,N}^{(\xi)(1)}] = 0$ and $\mathrm{E}_{\mathcal{D}}[\boldsymbol{q}_{i,N}^{(\xi)(2)}\boldsymbol{q}_{j,N}^{(\xi)(2)}] = 0$ for $i \neq j$. Next, we argue that

$$\sqrt{NT_1}\ \widetilde{\widetilde{\boldsymbol{m}}}_{(2)}^{(\xi)(2.1+2.5)} = \frac{1}{N_1\sqrt{N}}\sum_{i=2}^{N}\sum_{j<i}\boldsymbol{q}_{i,j}^{(\xi)} = o_P(1). \tag{A.35}$$

In particular, consider the variance of $\boldsymbol{q}_{i,N}^{(\xi)(1)}$ (the corresponding result for $\boldsymbol{q}_{i,N}^{(\xi)(2)}$ follows analogously):

$$\boldsymbol{\Sigma}_{\boldsymbol{q}^{(\xi)(1)},N} = \mathrm{E}_{\mathcal{D}}\left[\left(\frac{1}{N}\sum_{i=2}^{N}\boldsymbol{q}_{i,N}^{(\xi)(1)}\right)^2\right] = \frac{1}{N^2}\sum_{i=2}^{N}\mathrm{E}_{\mathcal{D}}\left[\left(\sum_{j<i}\boldsymbol{q}_{i,j}^{(\xi)(1)}\right)^2\right] = \frac{1}{N^2}\sum_{i=2}^{N}\sum_{j<i}\mathrm{E}_{\mathcal{D}}\left[\left(\boldsymbol{q}_{i,j}^{(\xi)(1)}\right)^2\right], \tag{A.36}$$

where the final equality holds by conditional independence. It remains to show that under our assumptions $\mathrm{E}_{\mathcal{D}}\left[\left(\boldsymbol{q}_{i,j}^{(\xi)(1)}\right)^2\right]$ are bounded by some constant Ξ. Observe that for all (i,j) we have

$$\boldsymbol{q}_{i,j}^{(\xi)(1)} = \frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}(\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1})\Delta g_{(q),j,t}^{(\xi)} = \frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\zeta}_{i,t}\psi_{j,t}^{(\xi)}; \tag{A.37}$$

$$\boldsymbol{q}_{i,j}^{(\xi)(2)} = \frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}(\boldsymbol{z}_{j,t}\Delta_f\varepsilon_{j,t+1})\Delta g_{(q),i,t}^{(\xi)} = \frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\zeta}_{j,t}\psi_{i,t}^{(\xi)}. \tag{A.38}$$

Here $\psi_{j,t}^{(\xi)} = \Delta g_{(q),j,t}^{(\xi)}$ by Assumption 3.1 satisfies $\mathrm{E}_{\mathcal{F}}[\psi_{j,t}^{(\xi)}] = 0$ and $\mathrm{E}_{\mathcal{D}}[|\psi_{j,t}^{(\xi)}|^2] < \Xi$. Moreover, $\boldsymbol{\zeta}_{i,t} = \boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t+1}$ satisfies $\mathrm{E}_{\mathcal{D}}[\boldsymbol{\zeta}_{i,t}] = 0$ and $\mathrm{E}_{\mathcal{D}}[|\boldsymbol{\zeta}_{i,t}|^{2+\delta}] < \Xi$. In addition, by assumption $\boldsymbol{\zeta}_{i,t}$ is a conditional mixing sequence with $\mu = 3(r+\delta)/\delta$ and $r = 4$. Thus, as all elements have a finite $4+\delta$ moment, by Lemma 2 it follows that $\mathrm{E}_{\mathcal{D}}\left[\left(\boldsymbol{q}_{i,j}^{(\xi)(1)}\right)^2\right] < \Xi$.

The remaining two components can be analyzed analogously. In particular:

$$\sqrt{NT_1}\ \widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)(2.3+2.7)} = \frac{1}{N_1\sqrt{NT_1}}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1}q_{j,t}\left(\varepsilon_{j,t+1}^{(\xi)}\boldsymbol{z}_{i,t}\varepsilon_{i,t} - \varepsilon_{j,t}^{(\xi)}\boldsymbol{z}_{i,t}\varepsilon_{i,t+1}\right) = o_P(1), \tag{A.39}$$

by conditional independence between (i,j) and the fact that all i and j random variables have zero expectations conditional on $\mathcal{D}$.

Finally, we consider $\widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)(2.4+2.8)}$. As previously, we define the following random variable $\Delta \boldsymbol{g}_{(\boldsymbol{z}),j,t} = \boldsymbol{z}_{j,t}\lambda_j - \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{j,t}\lambda_j]$, such that $\widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)(2.4+2.8)}$ can be expanded as follows:

$$\begin{aligned}\widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)(2.4+2.8)} &= \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1}\lambda_j \boldsymbol{z}_{j,t}q_{i,t}\Delta_f\varepsilon_{i,t+1}^{(\xi)}\\ &= \frac{1}{NT_1}\sum_{i=1}^{N}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_j]q_{i,t}\Delta_f\varepsilon_{i,t+1}^{(\xi)} + \frac{1}{NN_1T_1}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sum_{t=1}^{T_1}\Delta\boldsymbol{g}_{(\boldsymbol{z}),j,t}q_{i,t}\Delta_f\varepsilon_{i,t+1}^{(\xi)}\\ &= \widetilde{\widetilde{\boldsymbol{m}}}_{(1)}^{(\xi)(2.4+2.8)} + \widetilde{\widetilde{\boldsymbol{m}}}_{(2)}^{(\xi)(2.4+2.8)}.\end{aligned} \tag{A.40}$$

Since the second component has mean-zero, one can use identical steps to those used previously in establishing that $\widetilde{\widetilde{\boldsymbol{m}}}_{(2)}^{(\xi)(2.4+2.8)}$ is negligible. In particular, we have

$$\begin{aligned}\sqrt{NT_1}\ \widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)(2.4+2.8)} &= \frac{1}{\sqrt{NT_1}}\sum_{i=1}^{N}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_j]q_{i,t}\Delta_f\varepsilon_{i,t+1}^{(\xi)} + o_P(1)\\ &= \frac{1}{\sqrt{NT_1}}\sum_{i=1}^{N}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_j]q_{i,t}\Delta_f\varepsilon_{i,t+1}^{(\xi)} + o_P(1).\end{aligned} \tag{A.41}$$

Collecting all terms:

$$\begin{aligned}\sqrt{NT_1}\ \widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}_0) =& \frac{1}{\sqrt{NT_1}}\sum_{i=1}^{N}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^{(\xi)}]\boldsymbol{z}_{i,t}\left(f_{t+1}\varepsilon_{i,t} - f_t\varepsilon_{i,t+1}\right)\\ &- \frac{1}{\sqrt{NT_1}}\sum_{i=1}^{N}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]q_{i,t}\left(f_{t+1}\varepsilon_{i,t}^{(\xi)} - f_t\varepsilon_{i,t+1}^{(\xi)}\right) + o_P(1)\\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sqrt{T_1}\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)} + o_P(1).\end{aligned} \tag{A.42}$$

Next we establish $\mathcal{F}$ stable convergence of the leading term in $\sqrt{NT_1}\ \widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}_0)$. Denote by $\mathcal{D}_i$ the σ-field generated by $\sigma(\mathcal{F}\vee\{\boldsymbol{\Upsilon}_j\}_{j=1}^{i}\vee\{\mathcal{K}_j\}_{j=1}^{i})$, where $\boldsymbol{\Upsilon}_j$ and $\boldsymbol{v}_j$ are defined in Assumption 3.1. Then $\{\sqrt{T_1}\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}, \mathcal{D}_i : i \geq 1\}$ is a Martingale Difference sequence (element-wise), as by Assumption 3.1 all unit-specific variables are independent conditionally on $\mathcal{F}$. Also, let

$$\boldsymbol{\Omega}_T^{(\xi)} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[T_1\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}\left(\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}\right)'|\mathcal{D}_{i-1}\right] = \mathrm{E}_{\mathcal{F}}\left[T_1\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}\left(\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}\right)'\right], \tag{A.43}$$

and $\boldsymbol{\Omega}^{(\xi)} = \mathrm{plim}_{T\to\infty}\boldsymbol{\Omega}_T^{(\xi)}$. Using Theorem 3.2. and Corollary 3.1 in Hall and Heyde (1980) in conjunction with the Cramér-Wold device, yields

$$\sqrt{NT_1}\ \widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}_0) \Rightarrow \left(\boldsymbol{\Omega}^{(\xi)}\right)^{1/2}\boldsymbol{\psi}\quad\text{(stably)}, \tag{A.44}$$

where $\boldsymbol{\psi} \sim N(\boldsymbol{0}_D, \boldsymbol{I}_D)$. This result holds provided that each element $\overline{\mu}_{i,T}^{(\xi)(p)}$ for $p = 1,\ldots,D$ of $\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)}$ satisfies the conditional Lindeberg's condition:

$$N^{-1}\sum_{i=1}^{N}\mathrm{E}_{\mathcal{F}}[|\sqrt{T}_1\overline{\mu}_{i,T}^{(\xi)(p)}|^2 I(|\sqrt{T}_1\overline{\mu}_{i,T}^{(\xi)(p)}| > \sqrt{N}\varepsilon)],\quad \text{for all } \varepsilon > 0. \tag{A.45}$$

Given that the conditional Lyapunov's condition implies the conditional Lindeberg's condition, it is sufficient that $\mathrm{E}_{\mathcal{F}}[|\sqrt{T}_1\overline{\mu}_{i,T}^{(\xi)(p)}|^{2+\delta}] < \infty$ for some arbitrary $\delta > 0$. In our case this moment bound can be directly verified using Yokoyama's inequality for zero mean conditional mixing processes. Furthermore, by construction stable convergence implies $\mathcal{F}$-stable convergence.

Step 3. The proof of consistency is fairly standard along the lines of Newey and McFadden (1994). Sufficient conditions are satisfied based on uniform convergence and global identification over a compact set Θ by Assumption 3.3.

Step 4. Asymptotic distribution can be obtained by expanding the first-order conditions around the true value as in Eq. (34):

$$\boldsymbol{W}_{N,T}\widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{W}_{N,T}\left(\widetilde{\widetilde{\boldsymbol{m}}}^{(\xi)}(\boldsymbol{\beta}_0) + \left(\widetilde{\widetilde{\boldsymbol{\Gamma}}}^{(\xi)} + \frac{1}{2}\sum_{k=1}^{K}\widetilde{\widetilde{\boldsymbol{H}}}_k^{(\xi)}(\widehat{\beta}_k - \beta_{0,k})\right)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right). \tag{A.46}$$

This expansion is exact because the moment conditions are at most quadratic. Our assumptions ensure that $\text{plim}_{N,T\to\infty} \overline{\overline{\boldsymbol{\Gamma}}}^{(\xi)}$ and $\text{plim}_{N,T\to\infty} \overline{\overline{\boldsymbol{H}}}_k^{(\xi)}$, are finite and $\mathcal{F}$ measurable. Given that convergence in probability implies convergence in distribution, the remainder of the proof follows directly from Proposition A.2 in Kuersteiner and Prucha (2013) and an application of the continuous mapping theorem. □

A.5. Stacked moment conditions

Proof of Theorem 2. Analogous to the corresponding proof of Theorem 1, the result of this theorem can shown based on the following four distinct steps:

1. Analyze the bias term originating from "own" terms at $i = j$;
2. Derive the leading term of the asymptotic expansion;
3. Show consistency of the estimator;
4. Derive asymptotic distribution of the estimator.

As steps 3 and 4 are identical to those in Theorem 1, we mainly focus on steps 1 and 2. As in Theorem 1 we set $K = D = 1$, but continue using the vector and matrix notation unless it creates confusion. Finally, as the derivations for the nonlinear approach are identical to those in the linear approach, except for a larger number of negligible remainder terms, we provide the complete proof for the linear approach only.

Notice that the Jacobian matrix at time t can be decomposed as follows:

$$\overline{\boldsymbol{\Gamma}}_t^{(\xi)} = \boldsymbol{\Gamma}_t^{(\xi)} + \frac{1}{N}\boldsymbol{B}_t^{(\xi)} + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Z}_{i,t}^{(\xi)(1)} + \frac{1}{N^2}\sum_{i=1}^{N}\boldsymbol{Z}_{i,t}^{(\xi)(2)} + \boldsymbol{R}_{\boldsymbol{\Gamma},N,t}^{(\xi)}. \tag{A.47}$$

Here $\boldsymbol{\Gamma}_t^{(\xi)}$ is the expected Jacobian matrix as defined in Eq. (43), while $\boldsymbol{B}_t^{(\xi)}$ is a $\mathcal{F}$-measurable "bias" matrix. For the linear approach this is given by:

$$\begin{aligned}\boldsymbol{B}_t^{(L)} &= \mathrm{E}_{\mathcal{F}}\left[(q_{i,t}d_{i,t} - \mathrm{E}_{\mathcal{F}}[q_{i,t}d_{i,t}])(z_{i,t}\boldsymbol{x}_{i,t+1}' - \mathrm{E}_{\mathcal{F}}[z_{i,t}\boldsymbol{x}_{i,t+1}'])\right]\\ &\quad - \mathrm{E}_{\mathcal{F}}\left[(q_{i,t}d_{i,t+1} - \mathrm{E}_{\mathcal{F}}[q_{i,t}d_{i,t+1}])(z_{i,t}\boldsymbol{x}_{i,t}' - \mathrm{E}_{\mathcal{F}}[z_{i,t}\boldsymbol{x}_{i,t}'])\right].\end{aligned} \tag{A.48}$$

The influence functions $\boldsymbol{Z}_{i,t}^{(\xi)(1)}$ and $\boldsymbol{Z}_{i,t}^{(\xi)(2)}$ satisfy the conditional mean restriction of the form:

$$\mathrm{E}_{\mathcal{F}}\left[\boldsymbol{Z}_{i,t}^{(\xi)(1)}\right] = \boldsymbol{0}_{Z\times K}; \tag{A.49}$$

$$\mathrm{E}_{\mathcal{F}}\left[\boldsymbol{Z}_{i,t}^{(\xi)(2)}\right] = \boldsymbol{0}_{Z\times K}. \tag{A.50}$$

However, in general it is not true that the above conditions hold conditional on $\mathcal{D}$. Finally, the remainder term is of the form:

$$\boldsymbol{R}_{\boldsymbol{\Gamma},N,t}^{(\xi)} = \frac{1}{N^2}\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{i\neq j}^{N}\boldsymbol{v}_{i,t}^{(c)}(\boldsymbol{w}_{j,t}^{(c)})', \tag{A.51}$$

where $\mathrm{E}_{\mathcal{F}}[\boldsymbol{v}_{i,t}^{(c)}] = \boldsymbol{0}_D$ and $\mathrm{E}_{\mathcal{F}}[\boldsymbol{w}_{i,t}^{(c)}] = \boldsymbol{0}_K$ for all $i = 1,\dots,N$ and $t = 1,\dots,T$. The value of C depends on ξ; in particular, for $\xi = L$ we have $C = 2$, while for $\xi = NL$, $C = 4$. The exact form of $\boldsymbol{v}_{i,t}^{(c)}$ and $\boldsymbol{w}_{i,t}^{(c)}$ also depends on ξ, however these terms are asymptotically negligible and so they can be ignored.

Next, using the proof of Lemma 1 and Theorem 1 we can similarly expand vector of estimating equations:

$$\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0) = \frac{1}{N}\boldsymbol{b}_t^{(\xi)} + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\mu}_{i,t}^{(\xi)} + \frac{1}{N^2}\sum_{i=1}^{N}\boldsymbol{z}_{i,t}^{(\xi)(2)} + \boldsymbol{r}_{\boldsymbol{m},N,t}^{(\xi)}. \tag{A.52}$$

As it was shown in Theorem 1, the leading term $\boldsymbol{\mu}_{i,t}^{(\xi)}$, satisfies:

$$\mathrm{E}_{\mathcal{D}}\left[\boldsymbol{\mu}_{i,t}^{(\xi)}\right] = \boldsymbol{0}_D. \tag{A.53}$$

On the other hand, in general, for the other influence function only the condition $\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}^{(\xi)(2)}] = \boldsymbol{0}$ is satisfied. Finally, the remainder term is of the form:

$$\boldsymbol{r}_{\boldsymbol{m},N,t}^{(\xi)} = \frac{1}{N^2}\sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{i\neq j}^{N}\boldsymbol{o}_{i,t}^{(r)}u_{j,t}^{(r)}, \tag{A.54}$$

where for each $r = 1,\dots,R$ either $\mathrm{E}_{\mathcal{D}}[\boldsymbol{o}_{i,t}^{(c)}] = \boldsymbol{0}$ or $\mathrm{E}_{\mathcal{D}}[u_{i,t}^{(c)}] = \boldsymbol{0}$ or both. This is an important distinction between the remainder term of the estimating equations, and that of the Jacobian. In total, $R = 4$ irrespective of the approach considered.

The leading term in the asymptotic expansion of the first-order conditions is given by

$$\boldsymbol{l}_{N,T}^{(\xi)} = \frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}\right)'\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0). \tag{A.55}$$

This result is shown in Step 4.

We expand $\boldsymbol{l}_{N,T}^{(\xi)}$ into a sum of 20 distinct components as follows,

$$\boldsymbol{l}_{N,T}^{(\xi)} = \sum_{\ell_1=1}^{5}\sum_{\ell_2=1}^{4}\boldsymbol{l}_{N,T}^{(\xi)(\ell_1.\ell_2)}, \tag{A.56}$$

and analyze every term individually. Here we adopt the convention that ℓ_1 corresponds to the order of the element in Eq. (A.47), while ℓ_2 corresponds to Eq. (A.52).

The "numerator" of the incidental parameters bias term $\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)IP}$ is determined from:

$$\boldsymbol{l}_{N,T}^{(\xi)(1.1)} = \frac{1}{NT_1}\sum_{t=1}^{T_1}\left(\boldsymbol{\Gamma}_t^{(\xi)}\right)'\boldsymbol{s}_t = \frac{1}{N}\boldsymbol{s}_{T,\mathcal{F}}^{(\xi)IP}. \tag{A.57}$$

The leading variance term is given by:

$$\boldsymbol{l}_{N,T}^{(\xi)(1.2)} = \frac{1}{NT_1}\sum_{i=1}^{N}\sum_{t=1}^{T_1}\left(\boldsymbol{\Gamma}_t^{(\xi)}\right)'\boldsymbol{\mu}_{i,t}^{(\xi)} = \frac{1}{N}\sum_{i=1}^{N}\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)MMT}. \tag{A.58}$$

The "numerator" of the many-moments bias term $\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)MMT}$ is determined from the following component:

$$\begin{aligned}\boldsymbol{l}_{N,T}^{(\xi)(3.2)} &= \frac{1}{N}\frac{1}{NT_1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T_1}\left(\boldsymbol{Z}_{i,t}^{(\xi)(1)}\right)'\boldsymbol{\mu}_{j,t}^{(\xi)}\\ &= \frac{1}{N}\frac{1}{T_1}\sum_{t=1}^{T_1}\mathrm{E}_{\mathcal{F}}\left[\left(\boldsymbol{Z}_{i,t}^{(\xi)(1)}\right)'\boldsymbol{\mu}_{j,t}^{(\xi)}\right]\\ &+\frac{1}{N\sqrt{N}}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{T_1}\sum_{t=1}^{T_1}\left(\left(\boldsymbol{Z}_{i,t}^{(\xi)(1)}\right)'\boldsymbol{\mu}_{i,t}^{(\xi)}-\mathrm{E}_{\mathcal{F}}\left[\left(\boldsymbol{Z}_{i,t}^{(\xi)(1)}\right)'\boldsymbol{\mu}_{i,t}^{(\xi)}\right]\right)\\ &+\frac{1}{N}\frac{1}{NT_1}\sum_{i=1}^{N}\sum_{i\neq j}^{N}\sum_{t=1}^{T_1}\left(\boldsymbol{Z}_{i,t}^{(\xi)(1)}\right)'\boldsymbol{\mu}_{j,t}^{(\xi)}\\ &= \frac{1}{N}\boldsymbol{s}_{T,\mathcal{F}}^{(\xi)MMT} + o_P((NT_1)^{-1/2}).\end{aligned} \tag{A.59}$$

The final result follows upon appropriately defining $\boldsymbol{s}_{T,\mathcal{F}}^{(\xi)MMT}$. The negligibility of the second component is established using the Chebyshev's inequality, paired with the fact that $T/N \to \rho \in [0;\infty)$. The negligibility of the third component is established using the symmetrization argument used in Theorem 1. Notice that here we can use the symmetrization argument as $\mathrm{E}_{\mathcal{D}}[\boldsymbol{\mu}_{i,t}^{(\xi)}] = \boldsymbol{0}_D$.

It remains to show that all other 17 components are of order $o_P((NT_1)^{-1/2})$. The negligibility of the remaining components can be directly established based on one of the following approaches:

1. using the Markov's/Chebyshev's inequality for the terms denoted by superscripts (1.3), (2.1), (2.2), (2.3), (1.3), (3.3), (4.1), (4.2), (4.3), (5.1).
2. symmetrization as in Theorem 1 for the terms (1.4), (2.4).
3. Lemma 3 for the terms (3.4), (4.4), (5.2), (5.3), (5.4).

Combining everything we conclude that

$$\sqrt{NT_1}\left(\boldsymbol{l}_{N,T}^{(\xi)} - \frac{1}{N}\left(\boldsymbol{s}_{T,\mathcal{F}}^{(\xi)IP} + \boldsymbol{s}_{T,\mathcal{F}}^{(\xi)MMT}\right)\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sqrt{T_1}\overline{\boldsymbol{\mu}}_{i,T}^{(\xi)MMT} + o_P(1). \tag{A.60}$$

Stable convergence of the leading term can be established in a similar fashion as in Theorem 1.

Step 3. The proof of consistency is analogous to Theorem 1. Thus $\widehat{\boldsymbol{\beta}}_{MMT}^{(\xi)} - \boldsymbol{\beta}_0 = o_P(1)$.

Step 4. It remains to confirm that $\sqrt{NT_1}\left(\widehat{\boldsymbol{\beta}}_{MMT}^{(\xi)} - \boldsymbol{\beta}_0\right) = \mathcal{O}_P(1)$.

Analogous to Steps 1–2, one can show that for all $k = 1, \ldots, K$:

$$\sqrt{NT_1}\frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{H}}_{t,k}^{(\xi)}\right)'\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0) = \mathcal{O}_P(1). \tag{A.61}$$

The estimator $\widehat{\boldsymbol{\beta}}$ (we drop the subscript MMT, as well as the superscript (ξ) when the notation is unambiguous) solves the first order conditions:

$$\frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}(\widehat{\boldsymbol{\beta}})\right)'\overline{\boldsymbol{m}}_t^{(\xi)}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}_D, \tag{A.62}$$

where

$$\overline{\boldsymbol{\Gamma}}_t^{(\xi)}(\widehat{\boldsymbol{\beta}}) = \overline{\boldsymbol{\Gamma}}_t^{(\xi)} + \sum_{k=1}^{K}\overline{\boldsymbol{H}}_{t,k}^{(\xi)}(\widehat{\beta}_k - \beta_{0,k}). \tag{A.63}$$

Hence the first order conditions can be expanded as follows:

$$\frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}(\widehat{\boldsymbol{\beta}})\right)'\left(\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0) + \frac{1}{2}\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}(\widehat{\boldsymbol{\beta}}) + \overline{\boldsymbol{\Gamma}}_t^{(\xi)}\right)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right) = \mathbf{0}_D. \tag{A.64}$$

Re-arranging terms and using (A.63):

$$\begin{aligned}-\frac{1}{2}\left(\frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}(\widehat{\boldsymbol{\beta}})\right)'\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}(\widehat{\boldsymbol{\beta}}) + \overline{\boldsymbol{\Gamma}}_t^{(\xi)}\right)\right)\sqrt{NT_1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \frac{\sqrt{NT_1}}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}\right)'\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0)\\ &+ \sum_{k=1}^{K}(\widehat{\beta}_k - \beta_{0,k})\frac{\sqrt{NT_1}}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{H}}_{t,k}^{(\xi)}\right)'\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0).\end{aligned} \tag{A.65}$$

Notice that all products of the form

$$\frac{1}{T_1}\sum_{t=1}^{T_1}(\overline{\boldsymbol{\Gamma}}_t^{(\xi)})'(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}); \quad \frac{1}{T_1}\sum_{t=1}^{T_1}(\overline{\boldsymbol{\Gamma}}_t^{(\xi)})'(\overline{\boldsymbol{H}}_{t,k}^{(\xi)}); \quad \frac{1}{T_1}\sum_{t=1}^{T_1}(\overline{\boldsymbol{H}}_{t,k}^{(\xi)})'(\overline{\boldsymbol{H}}_{t,k}^{(\xi)}),$$

have well-defined probability limits by Assumption 3.6. This fact combined with consistency of the estimator implies that

$$-\left(\frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}\right)'\overline{\boldsymbol{\Gamma}}_t^{(\xi)} + o_P(1)\right)\sqrt{NT_1}(\widehat{\boldsymbol{\beta}}^{(\xi)} - \boldsymbol{\beta}_0) = \frac{\sqrt{NT_1}}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{\Gamma}}_t^{(\xi)}\right)'\overline{\boldsymbol{m}}_t^{(\xi)}(\boldsymbol{\beta}_0) + o_P(1). \tag{A.66}$$

From here the final result follows upon defining

$$\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)IP} = \left(\boldsymbol{\Gamma}_{MMT}^{(\xi)}\right)^{-1}\boldsymbol{s}_{T,\mathcal{F}}^{(\xi)IP}; \tag{A.67}$$

$$\boldsymbol{b}_{T,\mathcal{F}}^{(\xi)MMT} = \left(\boldsymbol{\Gamma}_{MMT}^{(\xi)}\right)^{-1}\boldsymbol{s}_{T,\mathcal{F}}^{(\xi)MMT}. \quad \square \tag{A.68}$$

Proof of Proposition 1. In terms of global identification, observe that

$$\begin{aligned}G^{(NL)}(\boldsymbol{\beta}) &= \operatorname*{plim}_{N,T\to\infty}\frac{1}{T_1}\sum_{t=1}^{T_1}\left(\overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta}) - \mathrm{E}[\overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta})]\right)'\left(\overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta}) - \mathrm{E}[\overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta})]\right)\\ &\quad + \operatorname*{plim}_{N\to\infty}\mathrm{E}[\overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta})]'\,\mathrm{E}[\overline{\boldsymbol{m}}_t^{(NL)}(\boldsymbol{\beta})]\\ &= G_{(1)}^{(NL)}(\boldsymbol{\beta}) + \boldsymbol{m}^{(NL)}(\boldsymbol{\beta})'\boldsymbol{m}^{(NL)}(\boldsymbol{\beta}).\end{aligned} \tag{A.69}$$

The conclusion follows by noting that $G^{(NL)}(\boldsymbol{\beta}) \geq \boldsymbol{m}^{(NL)}(\boldsymbol{\beta})'\boldsymbol{m}^{(NL)}(\boldsymbol{\beta})$, with an equality iff $G_{(1)}^{(NL)}(\boldsymbol{\beta}) = 0$. The proof for local identification is similar and therefore it is omitted. $\square$

A.6. Variance–covariance matrix estimation

Proof of Theorem 3. We split this proof into two parts. At first we prove the result for $\xi = L$, followed by $\xi = NL$. As in Theorem 1 we set $K = D = 1$, but continue using the vector and matrix notation unless it creates confusion.

Let $\Delta_f \boldsymbol{x}_{i,t+1} = f_t \boldsymbol{x}_{i,t+1} - f_{t+1}\boldsymbol{x}_{i,t}$. We use the Δ notation (without any subscript or superscript) to denote the deviations of the estimates from the corresponding true values, e.g. $\Delta\widehat{\varepsilon}_{i,s} = \widehat{\varepsilon}_{i,s} - \varepsilon_{i,s}$, $\Delta\widehat{\boldsymbol{g}}_{(\boldsymbol{z}),t,s} = \widehat{\boldsymbol{g}}_{(\boldsymbol{z}),t,s} - \boldsymbol{g}_{(\boldsymbol{z}),t,s}$, and $\Delta\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$.

For both approaches:

$$\Delta\widehat{\varepsilon}_{i,s} = \varepsilon_{i,s} + \lambda_i f_s - \boldsymbol{x}'_{i,s}\Delta\widehat{\boldsymbol{\beta}}; \tag{A.70}$$

$$\Delta\widehat{\boldsymbol{g}}_{(\boldsymbol{z}),t,s} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{z}_{i,t}\lambda_i - \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]\right)f_s + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_{i,t}\varepsilon_{i,s} - \left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_{i,t}\boldsymbol{x}'_{i,s}\right)\Delta\widehat{\boldsymbol{\beta}}. \tag{A.71}$$

Define $\Delta\boldsymbol{g}_{(\boldsymbol{z}),i,t} = \boldsymbol{z}_{i,t}\lambda_i - \mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]$, $\boldsymbol{Q}_{(\boldsymbol{zx}),i,t,s} = \boldsymbol{z}_{i,t}\boldsymbol{x}'_{i,s}$, $\boldsymbol{Q}_{(\boldsymbol{z}\varepsilon),i,t,s} = \boldsymbol{z}_{i,t}\varepsilon_{i,s}$, $\boldsymbol{Q}_{(\boldsymbol{z}\Delta\boldsymbol{x}),i,t,s} = \boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}'_{i,s}$, $\boldsymbol{Q}_{(\boldsymbol{z}\Delta\varepsilon),i,t,s} = \boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,s}$. Also let $\overline{\Delta\boldsymbol{g}}_{(\boldsymbol{z}),t}$, $\overline{\boldsymbol{Q}}_{(\boldsymbol{zx}),t,s}$, $\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),t,s}$ and $\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\Delta\varepsilon),t,s}$ define the corresponding cross-sectional averages.

Proof outline. Using the proof of Theorem 1 it is easy to show that:

$$\frac{1}{N\sqrt{T_1}}\sum_{i=1}^{N}\sum_{t=1}^{T_1}\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t} = o_P(1). \tag{A.72}$$

Thus, without any loss of generality, we prove this theorem for the un-centered version of the covariance matrix estimator:

$$\widehat{\boldsymbol{\Omega}}^{(\xi)}(\widehat{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t}\right)\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t}\right)'. \tag{A.73}$$

This expression can be conveniently expanded as

$$\begin{aligned}\widehat{\boldsymbol{\Omega}}^{(\xi)}(\widehat{\boldsymbol{\beta}}) &= \mathrm{E}_{\mathcal{F}}\left[\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\right)\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\right)'\right] \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\right)\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\right)' - \mathrm{E}_{\mathcal{F}}\left[\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\right)\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\right)'\right]\right] + \boldsymbol{R}_{\Omega} \\ &= \boldsymbol{\Omega}^{(\xi)} + o_P(1) + o_P(1) + \boldsymbol{R}_{\Omega}.\end{aligned} \tag{A.74}$$

Here the first $o_P(1)$ term is a by-product of Eq. (36), while the second $o_P(1)$ term follows from Chebyshev's inequality. The last term is of the form

$$\begin{aligned}\boldsymbol{R}_{\Omega} &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t}\right)\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t}\right)' + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\right)\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t}\right)' \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t}\right)\left(\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\right)'.\end{aligned} \tag{A.75}$$

Here $\Delta\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t}$ can be expressed as a sum of two distinct components $\Delta\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t} = (\widetilde{\boldsymbol{\mu}}^{(\xi)}_{i,t} - \boldsymbol{\mu}^{(\xi)}_{i,t}) + (\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t} - \widetilde{\boldsymbol{\mu}}^{(\xi)}_{i,t})$. The first component (denoted as $\Delta\widetilde{\boldsymbol{\mu}}^{(\xi)}_{i,t}$) is present purely due to the cross-sectional sampling uncertainty and is present even if $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$. The second component can be expanded as follows:

$$\widehat{\boldsymbol{\mu}}^{(\xi)}_{i,t} - \widetilde{\boldsymbol{\mu}}^{(\xi)}_{i,t} = \widetilde{\Delta}\boldsymbol{\Phi}^{(\xi)}_{i,t}\Delta\widehat{\boldsymbol{\beta}} + \sum_{k=1}^{K}\widetilde{\Delta}\boldsymbol{H}^{(\xi)}_{i,t,k}\Delta\widehat{\boldsymbol{\beta}}\Delta\widehat{\beta}_k. \tag{A.76}$$

Here for the linear approach we have $\widetilde{\Delta}\boldsymbol{H}^{(L)}_{i,t,k} = \boldsymbol{O}_{Z\times K}$,[31] as under our assumptions

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\boldsymbol{\mu}^{(\xi)}_{i,t}\| = \mathcal{O}(1). \tag{A.77}$$

To show that $\boldsymbol{R}_{\Omega} = o_P(1)$ it is sufficient to show that

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widetilde{\boldsymbol{\mu}}^{(\xi)}_{i,t}\|^2_{\infty} = o(1); \tag{A.78}$$

[31] Note that the implicit dependence on N is suppressed in the definition of all i specific variables.

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(\xi)}\|_{\max}^2 = o(NT_1); \tag{A.79}$$

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{H}_{i,t,k}^{(\xi)}\|_{\max}^2 = o((NT_1)^2). \tag{A.80}$$

These bounds paired with Markov's and Hölder's inequalities (element-wise) will deliver the desired result. Here the specific choice of vector and matrix norms is for notational convenience and is largely inconsequential for our results as all matrices are finite dimensional.

Linear approach: $\xi = L$. For the linear approach we also note that

$$\Delta\widehat{\varepsilon}_{i,s}^{d} = \lambda_i^d f_s; \tag{A.81}$$

$$\Delta\widehat{g}_{(q),t,s}^{(L)} = \frac{1}{N}\sum_{i=1}^{N}\left(q_{i,t}\lambda_i^d - \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d]\right) f_s + \frac{1}{N}\sum_{i=1}^{N} q_{i,t}\varepsilon_{i,s}^d. \tag{A.82}$$

Furthermore, let $\Delta g_{(q),i,t}^{(L)} = q_{i,t}\lambda_i^d - \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d]$, and define $\boldsymbol{Q}_{(q\varepsilon),i,t,s}^{(L)} = q_{i,t}\varepsilon_{i,s}^d$, $\boldsymbol{Q}_{(q\Delta\varepsilon),i,t,s}^{(L)} = q_{i,t}\Delta_f\varepsilon_{i,s}^d$, whereas $\overline{\Delta\boldsymbol{g}}_{(q),t}^{(L)}$, $\overline{\boldsymbol{Q}}_{(q\varepsilon),t,s}^{(L)}$ and $\overline{\boldsymbol{Q}}_{(q\Delta\varepsilon),t,s}^{(L)}$ denote the corresponding cross-sectional averages. At first we consider the deviation from the infeasible estimator:

$$\begin{aligned}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(L)} &= \boldsymbol{z}_{i,t}\lambda_i\overline{\boldsymbol{Q}}_{(q\Delta\varepsilon),t,t+1}^{(L)} - q_{i,t}\lambda_i^d\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\Delta\varepsilon)t,t+1} + \overline{\Delta\boldsymbol{g}}_{(\boldsymbol{z}),t}q_{i,t}\Delta_f\varepsilon_{i,t}^d - \overline{\Delta\boldsymbol{g}}_{(q),t}^{(L)}\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t}\\ &\quad + \overline{\boldsymbol{Q}}_{(q\varepsilon),t,t+1}^{(L)}\boldsymbol{z}_{i,t}\varepsilon_{i,t} - \overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),t,t+1}q_{i,t}\varepsilon_{i,t}^d + \overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),t,t}q_{i,t}\varepsilon_{i,t+1}^d - \overline{\boldsymbol{Q}}_{(q\varepsilon),t,t}^{(L)}\boldsymbol{z}_{i,t}\varepsilon_{i,t+1}\\ &= \sum_{s=1}^{8}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(L)(s)}.\end{aligned} \tag{A.83}$$

For the linear approach the matrix in Eq. (A.76) is of the form

$$\begin{aligned}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)} &= \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d]\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}' + \overline{\Delta\boldsymbol{g}}_{(q),t}^{(L)}\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}' + \overline{\boldsymbol{Q}}_{(\boldsymbol{z}\Delta\boldsymbol{x}),t,t+1}q_{i,t}\lambda_i^d\\ &\quad + \overline{\boldsymbol{Q}}_{(q\varepsilon),t,t+1}^{(L)}\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t}' - \overline{\boldsymbol{Q}}_{(q\varepsilon),t,t}^{(L)}\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t+1}' + \overline{\boldsymbol{Q}}_{(\boldsymbol{z}\boldsymbol{x}),t,t+1}q_{i,t}\varepsilon_{i,t}^d - \overline{\boldsymbol{Q}}_{(\boldsymbol{z}\boldsymbol{x}),t,t}q_{i,t}\varepsilon_{i,t+1}^d\\ &= \sum_{\ell=1}^{7}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(\ell)}.\end{aligned} \tag{A.84}$$

Given that $\|\Delta\widehat{\boldsymbol{\beta}}\| = \mathcal{O}_P((NT_1)^{-1/2})$ by Theorem 1 for the final result it is sufficient to show that for all s, ℓ:

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(L)(s)}\|_{\infty}^2 = o(1); \tag{A.85}$$

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(\ell)}\|_{\max}^2 = o(NT_1). \tag{A.86}$$

We will consider all components individually.

$$\begin{aligned}\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(L)(1)}\|_{\infty}^2 &= \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\sum_{j=1}^{N}\sum_{k=1}^{N}\mathrm{E}\left[\lambda_i^2\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,\tau}q_{j,t}\Delta_f\varepsilon_{j,t+1}^d q_{k,\tau}\Delta_f\varepsilon_{k,\tau+1}^d\right]\\ &= \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\mathrm{E}\left[\lambda_i^2\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,\tau}q_{i,t}\Delta_f\varepsilon_{i,t+1}^d q_{i,\tau}\Delta_f\varepsilon_{i,\tau+1}^d\right]\\ &\quad + \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\sum_{j=1}^{N}\mathrm{E}\left[\lambda_i^2\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,\tau}q_{j,t}\Delta_f\varepsilon_{j,t+1}^d q_{j,\tau}\Delta_f\varepsilon_{j,\tau+1}^d\right]\\ &= \mathcal{O}\left(T_1N^{-2}\right) + \mathcal{O}(N^{-1}).\end{aligned} \tag{A.87}$$

Here in the third line we used the Law of Iterated Expectations (LIE), conditional independence and the assumption that $\mathrm{E}_{\mathcal{D}}[\Delta_f\varepsilon_{j,t+1}^d q_{j,\tau}] = 0$. The final line follows from Assumption 3.1 with $r \geq 6$ (applied for the first component) and Lemma 2 (for the second component) as $\Delta_f\varepsilon_{j,t+1}^d q_{j,\tau}$ is a zero-mean conditional mixing sequence. The same idea can be used for $s = 2, 5, 6, 7, 8$ to show that the corresponding rates are $\mathcal{O}\left(T_1N^{-2}\right) + \mathcal{O}(N^{-1})$.

For the remaining two components the derivations need to be slightly modified.

$$\begin{aligned}
\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(L)(3)}\|_\infty^2 &= \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\sum_{j=1}^{N}\sum_{k=1}^{N}\mathrm{E}\left[q_{i,t}\Delta_f\varepsilon_{i,t+1}^d q_{i,\tau}\Delta_f\varepsilon_{i,\tau+1}^d\Delta\boldsymbol{g}_{(\boldsymbol{z}),j,t}\Delta\boldsymbol{g}_{(\boldsymbol{z}),k,\tau}\right]\\
&= \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\mathrm{E}\left[q_{i,t}\Delta_f\varepsilon_{i,t+1}^d q_{i,\tau}\Delta_f\varepsilon_{i,\tau+1}^d\Delta\boldsymbol{g}_{(\boldsymbol{z}),i,t}\Delta\boldsymbol{g}_{(\boldsymbol{z}),i,\tau}\right]\\
&+ \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\sum_{j=1}^{N}\mathrm{E}\left[q_{i,t}\Delta_f\varepsilon_{i,t+1}^d q_{i,\tau}\Delta_f\varepsilon_{i,\tau+1}^d\Delta\boldsymbol{g}_{(\boldsymbol{z}),j,t}\Delta\boldsymbol{g}_{(\boldsymbol{z}),j,\tau}\right]\\
&= \mathcal{O}\left(T_1N^{-2}\right)+\mathcal{O}(N^{-1}).
\end{aligned} \tag{A.88}$$

As with $s=1$, one can use the same arguments to establish the corresponding order of magnitude, but here $\Delta_f\varepsilon_{i,t+1}^d q_{i,\tau}$ is the zero-mean conditional mixing sequence necessary to apply the result of Lemma 2. Notice that for this result to hold, it is sufficient that $r=8$. It is not difficult to see that the same idea can be used for $s=4$. As a result,

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(L)(s)}\|_\infty^2 = \mathcal{O}\left(T_1N^{-2}\right)+\mathcal{O}(N^{-1}) = \mathrm{o}(1), \tag{A.89}$$

for all s, provided that $T_1N^{-2}\to 0$ as $N,T\to\infty$. It is easy to see that the same proof strategy can be used to show

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(\ell)}\|_{\max}^2 = \mathrm{o}(1), \quad \ell=4,5. \tag{A.90}$$

All other terms need to be analyzed differently. For example,

$$\begin{aligned}
\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(1)}\|_{\max}^2 &= \frac{1}{T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\mathrm{E}[\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i^d]\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}\,\mathrm{E}_{\mathcal{F}}[q_{i,\tau}\lambda_i^d]\boldsymbol{z}_{i,\tau}\Delta_f\boldsymbol{x}_{i,\tau+1}]\\
&= \mathcal{O}(T_1).
\end{aligned} \tag{A.91}$$

Moreover,

$$\begin{aligned}
\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(2)}\|_{\max}^2 &= \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\sum_{j=1}^{N}\sum_{k=1}^{N}\mathrm{E}\left[\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}\boldsymbol{z}_{i,\tau}\Delta_f\boldsymbol{x}_{i,\tau+1}\Delta g_{(q),j,t}\Delta g_{(q),k,\tau}\right]\\
&= \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\mathrm{E}\left[\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}\boldsymbol{z}_{i,\tau}\Delta_f\boldsymbol{x}_{i,\tau+1}\Delta g_{(q),i,t}\Delta g_{(q),i,\tau}\right]\\
&+ \frac{1}{N^2T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\sum_{j=1}^{N}\mathrm{E}\left[\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}\boldsymbol{z}_{i,\tau}\Delta_f\boldsymbol{x}_{i,\tau+1}\Delta g_{(q),j,t}\Delta g_{(q),j,\tau}\right]\\
&= \mathcal{O}\left(T_1N^{-2}\right)+\mathcal{O}(T_1N^{-1}).
\end{aligned} \tag{A.92}$$

Here in the third line we used the Law of Iterated Expectations (LIE), conditional independence and the assumption that $\mathrm{E}_{\mathcal{F}}[\Delta g_{(q),j,t}]=0$. The final line follows from Assumption 3.1 with $r\geq 6$ for the second component. Note that as $\mathrm{E}_{\mathcal{D}}[\Delta g_{(q),j,t}]\neq 0$, Lemma 2 does not apply, hence the $\mathcal{O}(T_1N^{-1})$ rate for the second component.

For $\ell=3$ note that

$$\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(3)} = \left(\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\Delta\boldsymbol{x}),t,t+1}-\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}]\right)q_{i,t}\lambda_i^d+\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}]q_{i,t}\lambda_i^d. \tag{A.93}$$

For the first part we can use the corresponding result from $\ell=1$, while for the second part the result from $\ell=2$. Thus

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(3)}\|_{\max}^2 = \mathcal{O}(T_1). \tag{A.94}$$

Finally, for $\ell=6$ consider a similar expansion:

$$\begin{aligned}
\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(6)} &= \left(\overline{\boldsymbol{Q}}_{(\boldsymbol{zx}),t,t+1}-\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t+1}]\right)q_{i,t}\varepsilon_{i,t}^d+\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t+1}]q_{i,t}\varepsilon_{i,t}^d\\
&= \widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(6.1)}+\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(6.2)}.
\end{aligned} \tag{A.95}$$

We have

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(6.2)}\|_{\max}^2 = \frac{1}{T_1}\sum_{t=1}^{T_1}\sum_{\tau=1}^{T_1}\mathrm{E}[\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t+1}]q_{i,t}\varepsilon_{i,t}^{d}\,\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,\tau}\boldsymbol{x}_{i,\tau+1}]q_{i,\tau}\varepsilon_{i,\tau}^{d}] = \mathcal{O}(1). \tag{A.96}$$

Here we use the fact that $\mathrm{E}_{\mathcal{D}}[\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t+1}]q_{i,t}\varepsilon_{i,t}^{d}]=0$ and that $q_{i,t}\varepsilon_{i,t}^{d}$ is a conditional mixing sequence. On the other hand, $\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(6.1)}$ can be analyzed analogously to $\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(L)(3)}$. Thus,

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)(6)}\|_{\max}^2 = \mathcal{O}(1)+\mathcal{O}(T_1N^{-2})+\mathcal{O}(N^{-2}). \tag{A.97}$$

For $\ell=7$ the derivations are equivalent. As a result, we conclude that

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(L)}\|_{\max}^2 = \mathcal{O}(T_1)=\mathrm{o}(NT_1), \tag{A.98}$$

as long as $N\to\infty$.

Notice that all T_1N^{-2} contributions originate from "own" terms due to the potential correlation between elements in the estimates $\widehat{\boldsymbol{g}}$ and all i specific random variables. Hence, if one uses individual specific estimates $\widehat{\boldsymbol{g}}_i$ then the result of this theorem follows without the additional restriction $T_1N^{-2}\to 0$.

Nonlinear approach: $\xi=NL$. For the nonlinear approach we note that some of the above derived results continue to hold upon noticing that $\varepsilon_{i,t}^{d}=\varepsilon_{i,t}$ and $\lambda_i^{d}=\lambda_i$. Moreover, at first we expand estimated quantities around the corresponding true values:

$$\Delta\widehat{\varepsilon}_{i,s}^{d} = \varepsilon_{i,s}+\lambda_i f_s-\boldsymbol{x}_{i,s}'\Delta\widehat{\boldsymbol{\beta}}; \tag{A.99}$$

$$\Delta\widehat{\boldsymbol{g}}_{(q),t,s} = \frac{1}{N}\sum_{i=1}^{N}\left(q_{i,t}\lambda_i-\mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i]\right)f_s+\frac{1}{N}\sum_{i=1}^{N}q_{i,t}\varepsilon_{i,s}-\left(\frac{1}{N}\sum_{i=1}^{N}q_{i,t}\boldsymbol{x}_{i,s}'\right)\Delta\widehat{\boldsymbol{\beta}}. \tag{A.100}$$

Hence, the deviation from the infeasible estimator is given by

$$\begin{aligned}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(NL)} &= \boldsymbol{z}_{i,t}\lambda_i\overline{\boldsymbol{Q}}_{(q\Delta\varepsilon),t,t+1}-q_{i,t}\lambda_i\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\Delta\varepsilon),t,t+1}+\overline{\Delta\boldsymbol{g}}_{(\boldsymbol{z})}q_{i,t}\Delta_f\varepsilon_{i,t}-\overline{\Delta g}_{(q)}\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t}\\ &\quad+\overline{\boldsymbol{Q}}_{(q\varepsilon),t,t+1}\boldsymbol{z}_{i,t}\varepsilon_{i,t}-\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),t,t+1}q_{i,t}\varepsilon_{i,t}+\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),t,t}q_{i,t}\varepsilon_{i,t+1}-\overline{\boldsymbol{Q}}_{(q\varepsilon),t,t}\boldsymbol{z}_{i,t}\varepsilon_{i,t+1}\\ &= \boldsymbol{z}_{i,t}\lambda_i\overline{\boldsymbol{Q}}_{(q\Delta\varepsilon),i,t,t+1}-q_{i,t}\lambda_i\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\Delta\varepsilon),i,t,t+1}+\overline{\Delta\boldsymbol{g}}_{(\boldsymbol{z}),i}q_{i,t}\Delta_f\varepsilon_{i,t}-\overline{\Delta g}_{(q),i}\boldsymbol{z}_{i,t}\Delta_f\varepsilon_{i,t}\\ &\quad+\overline{\boldsymbol{Q}}_{(q\varepsilon),i,t,t+1}\boldsymbol{z}_{i,t}\varepsilon_{i,t}-\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),i,t,t+1}q_{i,t}\varepsilon_{i,t}+\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),i,t,t}q_{i,t}\varepsilon_{i,t+1}-\overline{\boldsymbol{Q}}_{(q\varepsilon),i,t,t}\boldsymbol{z}_{i,t}\varepsilon_{i,t+1}\\ &= \sum_{s=1}^{8}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(NL)(s)}.\end{aligned} \tag{A.101}$$

Here we use subscript i on cross-sectional averages to denote the corresponding delete-one versions, e.g.

$$\overline{\boldsymbol{Q}}_{(q\Delta\varepsilon),i,t,t+1} = \frac{1}{N}\sum_{j\neq i}q_{j,t}\Delta_f\varepsilon_{j,t+1}, \tag{A.102}$$

and so on. Because of the automatic correction for own terms, it is easy to show that

$$\mathrm{E}\,\|\frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\Delta\widetilde{\boldsymbol{\mu}}_{i,t}^{(NL)(s)}\|_{\infty}^2 = \mathcal{O}(N^{-1})=\mathrm{o}(1), \tag{A.103}$$

for all s. As a result, all contributions are asymptotically negligible without any restrictions on relative rates of N and T.

The first-order contribution in Eq. (A.76) is of the form

$$\begin{aligned}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(NL)} &= \mathrm{E}_{\mathcal{F}}[q_{i,t}\lambda_i]\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}'+\overline{\Delta g}_{(q),t}\boldsymbol{z}_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}'+q_{i,t}\lambda_i\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\Delta\boldsymbol{x}),t,t+1}\\ &\quad+\overline{\boldsymbol{Q}}_{(q\varepsilon),t,t+1}\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t}'-\overline{\boldsymbol{Q}}_{(q\varepsilon),t,t}\boldsymbol{z}_{i,t}\boldsymbol{x}_{i,t+1}'+\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\boldsymbol{x}),t,t+1}q_{i,t}\varepsilon_{i,t}-\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\boldsymbol{x}),t,t}q_{i,t}\varepsilon_{i,t+1}\\ &\quad-\mathrm{E}_{\mathcal{F}}[\boldsymbol{z}_{i,t}\lambda_i]q_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}'-\overline{\Delta\boldsymbol{g}}_{(\boldsymbol{z}),t}q_{i,t}\Delta_f\boldsymbol{x}_{i,t+1}'-\boldsymbol{z}_{i,t}\lambda_i\overline{\boldsymbol{Q}}_{(q\Delta\boldsymbol{x}),t,t+1}\\ &\quad-\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),t,t+1}q_{i,t}\boldsymbol{x}_{i,t}'+\overline{\boldsymbol{Q}}_{(\boldsymbol{z}\varepsilon),t,t}q_{i,t}\boldsymbol{x}_{i,t+1}'-\boldsymbol{z}_{i,t}\overline{\boldsymbol{Q}}_{(q\boldsymbol{x}),t,t+1}\varepsilon_{i,t}+\boldsymbol{z}_{i,t}\overline{\boldsymbol{Q}}_{(q\boldsymbol{x}),t,t}\varepsilon_{i,t+1}\\ &= \sum_{\ell=1}^{14}\widetilde{\Delta}\boldsymbol{\Phi}_{i,t}^{(NL)(\ell)}.\end{aligned} \tag{A.104}$$

Finally, consider the second-order contribution in Eq. (A.76):

$$\begin{aligned}\widetilde{\Delta}\boldsymbol{H}^{(NL)}_{i,t,k} &= \boldsymbol{z}_{i,t}x^{(k)}_{i,t}\overline{\mathbf{Q}}_{(q\boldsymbol{x}),t,t+1} - \boldsymbol{z}_{i,t}x^{(k)}_{i,t+1}\overline{\mathbf{Q}}_{(q\boldsymbol{x}),t,t} - q_{i,t}x^{(k)}_{i,t}\overline{\mathbf{Q}}_{(\boldsymbol{zx}),t,t+1} + q_{i,t}x^{(k)}_{i,t+1}\overline{\mathbf{Q}}_{(\boldsymbol{zx}),t,t} \\ &= \boldsymbol{z}_{i,t}x^{(k)}_{i,t}\overline{\mathbf{Q}}_{(q\boldsymbol{x}),i,t,t+1} - \boldsymbol{z}_{i,t}x^{(k)}_{i,t+1}\overline{\mathbf{Q}}_{(q\boldsymbol{x}),i,t,t} - q_{i,t}x^{(k)}_{i,t}\overline{\mathbf{Q}}_{(\boldsymbol{zx}),i,t,t+1} + q_{i,t}x^{(k)}_{i,t+1}\overline{\mathbf{Q}}_{(\boldsymbol{zx}),i,t,t} \\ &= \sum_{s=1}^{4}\widetilde{\Delta}\boldsymbol{H}^{(NL)(s)}_{i,t,k}. \end{aligned} \tag{A.105}$$

Using steps similar to those used to bound $\widetilde{\Delta}\boldsymbol{\Phi}^{(L)(s)}_{i,t}$, one can show that

$$\mathrm{E}\,\| \frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{\Phi}^{(NL)}_{i,t}\|^2_{\max} = \mathcal{O}(T_1) = \mathrm{o}(NT_1); \tag{A.106}$$

$$\mathrm{E}\,\| \frac{1}{\sqrt{T_1}}\sum_{t=1}^{T_1}\widetilde{\Delta}\boldsymbol{H}^{(NL)}_{i,t,k}\|^2_{\max} = \mathcal{O}(T_1) = \mathrm{o}(NT_1). \quad \square \tag{A.107}$$

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.03.011.

References

Ahn, S., Lee, Y.H., Schmidt, P., 2001. GMM Estimation of linear panel data models with time-varying individual effects. J. Econometrics 101 (1), 219–255.
Ahn, S., Lee, Y.H., Schmidt, P., 2013. Panel data models with multiple time-varying individual effects. J. Econometrics 174 (1), 1–14.
Ahn, S., Schmidt, P., 1995. Efficient estimation of models for dynamic panel data. J. Econometrics 68, 5–27.
Alvarez, J., Arellano, M., 2003. The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica 71 (4), 1121–1159.
Anderson, T., Hsiao, C., 1982. Formulation and estimation of dynamic models using panel data. J. Econometrics 18, 47–82.
Angrist, J.D., Imbens, G.W., Krueger, A.B., 1999. Jackknife instrumental variables estimation. J. Appl. Econometrics 14 (1), 57–67.
Angrist, J.D., Krueger, A.B., 1995. Split-sample instrumental variables estimates of the return to schooling. J. Bus. Econom. Statist. 13 (2), 225–235.
Antoine, B., Renault, E., 2009. Efficient GMM with nearly-weak instruments. Econom. J. 12 (s1), S135–S171.
Arellano, M., 1989. A note on the Anderson-hsiao estimator for panel data. Econom. Lett. 31 (4), 337–341.
Arellano, M., 2003. Panel Data Econometrics. In: Advanced Texts in Econometrics, Oxford University Press.
Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77 (4), 1229–1279.
Bai, J., 2013. Likelihood Approach to Dynamic Panel Models with Interactive Effects. mimeo.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.
Bekker, P.A., 1994. Alternative approximations to the distributions of instrumental variable estimators. Econometrica 62 (3), 657–681.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21 (1), C19–C68.
Chudik, A., Pesaran, M.H., 2015a. Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. J. Econometrics 188 (2), 393–420.
Chudik, A., Pesaran, M.H., 2015b. Large panel data models with cross-sectional dependence: A survey. In: Baltagi, B.H. (Ed.), The Oxford Handbook of Panel Data. Oxford University Press, Oxford, (Chapter 1).
Cui, G., Norkutė, M., Sarafidis, V., Yamagata, T., 2020. Two-Step Instrumental Variable Estimation of Linear Panel Data Models with Interactive Effects. Working Paper, Monash University.
Dhaene, G., Jochmans, K., 2015. Split-panel jackknife estimation of fixed-effect models. Rev. Econom. Stud. 82 (3), 991–1030.
Doukhan, P., 1994. Mixing: Properties and Examples. In: Lecture Notes in Statistics, vol. 85, Springer-Verlag.
Everaert, G., Pozzi, L., 2014. The predictability of aggregate consumption growth in OECD countries: A panel data analysis. J. Appl. Econometrics 29 (3), 431–453.
Fan, J., Liao, Y., 2020. Learning latent factors from diversified projections and its applications to over-estimated and weak factors. J. Amer. Statist. Assoc. (ja), 1–42.
Fernández-Val, I., Freeman, H., Weidner, M., 2021. Low-Rank Approximations of Nonseparable Panel Models. Econom. J. (forthcoming).
Fernández-Val, I., Weidner, M., 2016. Individual and time effects in nonlinear panel models with large n, t. J. Econometrics 192 (1), 291–312.
Galvao, A.F., Kato, K., 2014. Estimation and inference for linear panel data models under misspecification when both n and t are large. J. Bus. Econom. Statist. 32 (2), 285–309.
Geweke, J., Meese, R., 1981. Estimating regression models of finite but unknown order. J. Econometrics 16 (1), 162.
Graham, B., 2017. An econometric model of network formation with degree heterogeneity. Econometrica 85 (4), 1033–1063.
Hahn, J., Hausman, J., Kuersteiner, G., 2007. Long difference instrumental variables estimation for dynamic panel models with fixed effects. J. Econometrics 140 (2), 574–617.
Hall, P., Heyde, C.C., 1980. Martingale Limit Theory and Its Application. In: Probability and Mathematical Statistics, Academic Press.
Han, C., Phillips, P.C.B., 2006. GMM With many moment conditions. Econometrica 74 (1), 147–192.
Hansen, C.B., 2007. Asymptotic properties of a robust variance matrix estimator for panel data when t is large. J. Econometrics 141 (2), 597–620.
Hansen, C., Liao, Y., 2018. The factor-lasso and K-step bootstrap approach for inference in high-dimensional economic applications. Econometric Theory 35, 465–509.
Hayakawa, K., 2011. GMM Estimation of short dynamic panel data model with interactive fixed effects. J. Japan Statist. Soc. 42, 109–123.
Hayakawa, K., 2016. Identification problem of GMM estimators for short panel data models with interactive fixed effects. Econom. Lett. 139, 22–26.
Holtz-Eakin, D., Newey, W.K., Rosen, H.S., 1988. Estimating vector autoregressions with panel data. Econometrica 56, 1371–1395.
Hong, S., Su, L., Wang, Y., 2019. Inference in Partially Identified Panel Data Models with Interactive Fixed Effects. mimeo.
Jochmans, K., 2017. Two-way models for gravity. Rev. Econ. Stat. 99 (3), 478–485.

Juodis, A., 2018. Pseudo panel data models with cohort interactive effects. J. Bus. Econom. Statist. 36 (1), 47—61.
Juodis, A., 2020a. A Regularization Approach to Common Correlated Effects Estimation. mimeo.
Juodis, A., 2020b. This Shock is Different: Estimation and Inference in Misspecified Two Way Fixed Effects Regressions. mimeo.
Juodis, A., Karabiyik, H., Westerlund, J., 2021. On the robustness of the pooled CCE estimator. J. Econometrics 220 (2), 325–348.
Juodis, A., Sarafidis, V., 2018. Fixed t dynamic panel data estimators with multi-factor errors. Econometric Rev. 37 (8), 893–929.
Juodis, A., Sarafidis, V., 2020. A linear estimator for factor-augmented fixed-t panels with endogenous regressors. J. Bus. Econom. Statist. (forthcoming).
Kapetanios, G., 2008. A bootstrap procedure for panel data sets with many cross-sectional units. Econom. J. 11 (2), 377–395.
Kuersteiner, G., Prucha, I.R., 2013. Limit theory for panel data models with cross sectional dependence and sequential exogeneity. J. Econometrics 174 (2), 107–126.
Kuersteiner, G., Prucha, I.R., 2020. Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity. Econometrica 88 (5), 2109–2146.
Lee, N., Moon, H.R., Weidner, M., 2012. Analysis of interactive fixed effects dynamic linear panel regression with measurement error. Econom. Lett. 117 (1), 239–242.
Lee, N., Moon, H.R., Zhou, Q., 2017. Many IVs estimation of dynamic panel regression models with measurement error. J. Econometrics 200 (2), 251–259, Measurement Error Models.
Lu, X., Su, L., 2016. Shrinkage estimation of dynamic panel data models with interactive fixed effects. J. Econometrics 190 (1), 148–175.
Menzel, K., 2019. Bootstrap with Cluster-Dependence in Two or More Dimensions. Working Paper.
Moon, H.R., Weidner, M., 2017. Dynamic linear panel regression models with interactive fixed effects. Econometric Theory 33 (1), 158–195.
Nauges, C., Thomas, A., 2003. Consistent estimation of dynamic panel data models with time-varying individual effects. Ann. Econ. Statist. 70, 54–75.
Newey, W., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, Vol. 4. Elsevier, Amsterdam, pp. 2111–2245.
Newey, W., Windmeijer, F., 2009. GMM With many weak moment conditions. Econometrica 77, 687–719.
Norkutė, M., Sarafidis, V., Yamagata, T., Cui, G., 2021. Instrumental variable estimation of dynamic linear panel data models with defactored regressors and a multifactor error structure. J. Econometrics 220 (2), 416–446.
Pesaran, M.H., 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74 (4), 967–1012.
Pesaran, M.H., Smith, L.V., Yamagata, T., 2013. Panel unit root tests in the presence of a multifactor error structure. J. Econometrics 175 (2), 94–115.
Qian, J., Su, L., 2016. Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso. J. Econometrics 191 (1), 86–109.
Robertson, D., Sarafidis, V., 2015. IV Estimation of panels with factor residuals. J. Econometrics 185 (2), 526–541.
Robertson, D., Sarafidis, V., Westerlund, J., 2018. Unit root inference in generally trending and cross-correlated dynamic panels. J. Bus. Econom. Statist. 36 (3), 493–504.
Sarafidis, V., Wansbeek, T., 2021. Celebrating 40 years of panel data analysis: Past, present and future. J. Econometrics 220 (2), 215–226.
Satchachai, P., Schmidt, P., 2008. GMM With more moment conditions than observations. Econom. Lett. 99, 252–255.
Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586.
Su, L., Jin, S., 2012. Specification test for panel data models with interactive fixed effects. J. Econometrics 169 (1), 34–47.
Su, L., Jin, S., Zhang, Y., 2015. Specification test for panel data models with interactive fixed effects. J. Econometrics 186 (1), 222–244.
van der Vaart, A., 2000. Asymptotic Statistics. In: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.
Westerlund, J., Petrova, Y., Norkutė, M., 2019. CCE In fixed-t panels. J. Appl. Econometrics 34 (5), 746–761.