On the robustness of the pooled CCE estimator

Juodis, A.; Karabiyik, H.; Westerlund, J.

DOI
10.1016/j.jeconom.2020.06.002
Publication date
2021
Document Version
Final published version
Published in
Journal of Econometrics

Journal of Econometrics 220 (2021) 325–348

Contents lists available at ScienceDirect

Journal of Econometrics

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

On the robustness of the pooled CCE estimator☆

Artūras Juodis [a,*], Hande Karabiyik [b], Joakim Westerlund [c,d]

[a] Amsterdam School of Economics, University of Amsterdam, The Netherlands
[b] School of Business and Economics, Vrije Universiteit Amsterdam, The Netherlands
[c] Department of Economics, Lund University, Sweden
[d] Center for Financial Econometrics, Deakin University, Australia

ARTICLE INFO

Article history:
Available online 20 June 2020

JEL classification:
C13
C15
C23

Keywords:
Factor-augmented regressions
Rank condition
Common correlated effects
Predetermined regressors

ABSTRACT

Among the existing estimators of factor-augmented regressions, the CCE approach is the most popular. A major reason for this popularity is the simplicity and good small-sample performance of the approach, making it very attractive from an empirical point of view. The main drawback is that most of the available asymptotic theory is based on quite restrictive assumptions, such as that the common factor component should be independent of the regressors. The present paper can be seen as a reaction to this. The purpose is to study the asymptotic properties of the pooled CCE estimator under more realistic conditions. In particular, the common factor component may be correlated with the regressors, and the true number of common factors, r, can be larger than the number of estimated factors, which in CCE is given by $k+1$, where k is the number of regressors. The main conclusion is that while the estimator is generally consistent, asymptotic normality can fail when $r > k+1$.

1. Introduction

Panel data models typically include unit- and time-specific fixed effects to account for unobserved characteristics. While in most models these effects enter additively, in this paper they are interacted multiplicatively, leading to a factor-augmented regression model with the time-specific effects as factors and the unit-specific effects as factor loadings. This common factor specification contains the conventional fixed effects models as special cases, but is much more flexible since it allows the factors to affect each cross-section unit differently.

One of the most popular estimation approaches to factor-augmented regression models is the common correlated effects (CCE) approach of Pesaran (2006), which is based on taking the cross-sectional averages of the observables as estimated factors, and applying ordinary least squares (OLS) with the estimated factors in place of their true counterparts. Two of the reasons for the popularity of this approach are its simplicity and good small-sample performance when

☆ We would like to thank Vasilis Sarafidis (Guest Co-Editor), two anonymous referees and the participants of the 2017 International Panel Data Conference in Thessaloniki, the 2017 Bristol Econometric Study Group Meeting, the Bank of Lithuania, and the LMU (Munich) for helpful comments. Part of this paper was written while the first author enjoyed the hospitality of the Department of Economics at Lund University in Spring 2015 with the financial support of the C. Willems Stichting. The financial support via the NWO MaGW (number 404–10–457), The Netherlands and NWO VENI (number 451–17–002) grants is gratefully appreciated by the first author. The third author would like to thank the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship, Sweden, and the Jan Wallander and Tom Hedelius Foundation, Sweden for financial support under research Grant Number P2014–0112:1.
* Correspondence to: Amsterdam School of Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands.
E-mail address: a.juodis@uva.nl (A. Juodis).

https://doi.org/10.1016/j.jeconom.2020.06.002
0304-4076/

compared to the main competitor based on quasi maximum-likelihood (QML) (see, for example, Chudik et al., 2011; Chudik and Pesaran, 2015, and Westerlund and Urbain, 2015).

But while simple and with good small-sample performance, the data generating processes (DGPs) considered in the CCE strand of the literature are often less appealing than those considered in the QML strand. One of the reasons for this is that the loadings are typically assumed to be random coefficients that are independent of all other random elements of the model, including the regressors, which means that, in analogy with the literature on omitted variables, the factors can be ignored. Hence, for consistency there is no need for any augmentation by estimated factors in the first place. In other words, the independent loadings condition is not only unlikely to be met in practice, but it also leads to a rather simplified estimation problem. Moreover, the rate of convergence is reduced, from the usual $\sqrt{NT}$ to $\sqrt{N}$, which means that the conventional bias terms due to the incidental parameters are asymptotically negligible. In simulations, however, there is a clear bias effect (see, for example, Chudik and Pesaran, 2015), suggesting that the asymptotic approximation that ignores incidental parameter biases can be poor.

Another issue is that unlike in most existing QML studies where the regressors are given a nonparametric treatment, in CCE the regressors have to admit to a common factor representation that features the same set of factors as the dependent variable, which is questionable (see, for example, Moon and Weidner, 2017).

In this paper we try to combine the best of two worlds; we take the simplicity and good small-sample performance of CCE and apply it to a DGP that is similar to those considered in the QML literature. The DGP is therefore very general and includes many existing models as special cases.[1] In particular, unlike most existing CCE studies, the regressors are not required to be exogenous but may be predetermined (see Chudik and Pesaran, 2015, and Everaert and De Groote, 2016 for two exceptions). The type of factors that can be accommodated is also very general. We allow for r factors, which may be smaller or larger than the number of observables, $k+1$, provided that $k+1 \geq r_m$, where r_m is the number of "mean factors". The main difference here is, as the name suggests, that while the mean factors affect the mean of the data, the non-mean factors do not. This is important, because the cross-sectional averages of the observables will only be able to capture those factors that affect the mean. Hence, we essentially allow some factors to be inestimable when using the cross-sectional averages, which is of course always a risk in practice. Moreover, the non-mean factors are not restricted to be strong, but could also be weak as in Chudik et al. (2011), which means that the cross-sectional dependence can be spatial in nature. The distinction between mean and non-mean factors stands in sharp contrast to the existing CCE literature based on equal slopes, which supposes not only that all r factors are estimable, but also that $r \leq k+1$, a condition that is again not necessary in the present study.

As already mentioned, the CCE assumption that both the dependent variable and the regressors load on the same set of factors is restrictive. In the present paper, we allow $r_y \leq r_m$ of the mean factors, henceforth referred to as "y-factors", to enter the equation for the dependent variable, while the remaining factors enter only indirectly through the regressors. The factors that enter the equation for the regressors are therefore not necessarily the same as those that enter the equation for the dependent variable. The main restriction is that the y-factors must be estimable based on the cross-sectional averages, which is analogous to the QML literature.

There are two types of CCE estimators; there is the mean group CCE (CCEMG) estimator, which is obtained by taking the cross-section average of individual time series CCE estimators, and there is the pooled CCE (CCEP) estimator. Most results in the literature are for the former estimator, which can be difficult to analyze, because of its construction as a sum of ratios of random variables. The main exception is when the loadings are independent of the regressors, in which the asymptotic analysis simplifies quite substantially. In this study, we again follow the QML strand of the literature and focus on the CCEP estimator, which has not only certain optimality properties, but also relatively good performance in simulations (see, for example, Pesaran, 2006; Kapetanios et al., 2011, and Westerlund and Urbain, 2015). The question we ask is: Under what conditions will the CCEP estimator be consistent and asymptotically normal? To answer this, we begin by deriving an asymptotic expansion of the estimator. The expansion is very general and retains all terms that are non-negligible when the rate of convergence is $\sqrt{NT}$. The assumption we make is that $N/T \to \kappa \in (0, \infty)$ as $N, T \to \infty$ jointly, which is similar to the assumptions used in, for example, Bai (2009), Moon and Weidner (2015, 2017), Karabiyik et al. (2017), and Westerlund and Urbain (2015). As a natural second step in our analysis, the asymptotic expansion is used to evaluate the consistency and asymptotic distribution of the CCEP estimator. In so doing, it turns out to be useful to consider a few different specifications of k, r_m and r. The results show that $\sqrt{NT}$-consistency and asymptotic normality is generally possible when either $k+1 = r_m$, so that the number of estimated factors is equal to the number of mean factors, or $r_m = r$, so that all the factors are estimable using CCE. If, however, $r_m < \min\{r, k+1\}$, asymptotic normality generally breaks down, although consistency is still there. The rate of convergence in this last case is generally given by $\sqrt{N}$, although we also list conditions that ensure that the rate is $\sqrt{NT}$.

In this paper we assume that both dimensions (cross-sectional and time-series) of the data are large. If the length of time-series is limited (i.e. T fixed), then the results obtained in this paper are not directly applicable. For estimation

[1] The main restriction is that the slopes of the regressors must be equal across the cross-section, which is the same as in, for example, Bai (2009), Greenaway-McGrevy et al. (2012), and Moon and Weidner (2015, 2017). As usual with panel data models, we do not consider them unless there is some similarity that can be exploited. The idea here is to selectively pool the information regarding the slopes, and to impose only minor restrictions on the common factor component of the model, which will enable us to study the factor estimation problem when it matters, as it is commonly done in the QML literature (see, for example, Bai, 2009, and Moon and Weidner, 2015, 2017).

procedures applicable in such situations please refer to e.g. Holtz-Eakin et al. (1988), Ahn et al. (2013), Robertson and Sarafidis (2015), and Juodis and Sarafidis (2018).

The rest of the paper is organized as follows. Section 2 introduces the DGP, the CCEP estimator (Section 2.1), and the various specifications of k, r_m and r that we will be considering (Section 2.2). Because of the generality of the DGP, a large part of the section will be devoted to discussion. The main results are reported in Section 3, which are divided into asymptotic expansion (Section 3.1), consistency (Section 3.2), asymptotic distribution (Section 3.3), and implications for inference (Section 3.4). Section 4 reports the results of a small-scale Monte Carlo study. Section 5 concludes. All proofs are provided in the supplementary online appendix.

2. Model, estimator and specifications

2.1. Model and estimator

Consider the scalar panel data variable $y_{i,t}$, observable for $t = 1, \dots, T$ time series and $i = 1, \dots, N$ cross-sectional units. The DGP of the $T \times 1$ vector $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,T})'$ is given by

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{F}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i, \tag{2.1}$$

where $\mathbf{x}_i = (\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,T})'$ is a $T \times k$ matrix of regressors, $\mathbf{F} = (\mathbf{F}_1, \dots, \mathbf{F}_T)'$ is a $T \times r$ matrix of unobserved common factors, $\boldsymbol{\lambda}_i$ is a $r \times 1$ vector of factor loadings, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \dots, \varepsilon_{i,T})'$ is a $T \times 1$ vector of idiosyncratic errors.[2] If the composite error term $\mathbf{F}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$ is uncorrelated with $\mathbf{x}_i$, then (2.1) is nothing but a static panel data regression with exogenous regressors, which can be estimated consistently using OLS. If, however, $\mathbf{x}_i$ is correlated with $\mathbf{F}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$, then consistency will be lost. To allow for this possibility, we follow the existing CCE literature, and assume that

$$\mathbf{x}_i = \mathbf{F}\boldsymbol{\Lambda}_i' + \mathbf{v}_i, \tag{2.2}$$

where $\boldsymbol{\Lambda}_i$ is a $k \times r$ loading matrix and $\mathbf{v}_i = (\mathbf{v}_{i,1}, \dots, \mathbf{v}_{i,T})'$ is a $T \times k$ matrix of idiosyncratic errors. However, in contrast to most existing studies, which assume that $\mathbf{v}_i$ and $\boldsymbol{\varepsilon}_i$ are independent, in the present study $\mathbf{x}_i$ may be predetermined, such that $\boldsymbol{\varepsilon}_i$ is uncorrelated with current and past values of $\mathbf{x}_{i,t}$ but not with its future values. In particular, $\mathbf{x}_{i,t}$ may contain lags of $y_{i,t}$.

Example 2.1 (*The AR(1) Model*)**.** Consider the following first-order autoregressive, or AR(1), model with a single factor:

$$y_{i,t} = \alpha y_{i,t-1} + \gamma_i f_t + \varepsilon_{i,t}. \tag{2.3}$$

This model has attracted considerable interest in the literature (see, for example, Everaert and De Groote, 2016; Pesaran, 2007, and Westerlund, 2015), and is a special case of our more general DGP. In order to appreciate this, note how

$$y_{i,t} = \gamma_i(1 - \alpha L)^{-1} f_t + (1 - \alpha L)^{-1}\varepsilon_{i,t} = \gamma_i\alpha(L)f_t + \alpha(L)\varepsilon_{i,t}, \tag{2.4}$$

where L is the lag operator and $a(L) = (1 - aL)^{-1}$ for any constant $|a| < 1$. If we lag this last equation, we have a static factor model for the regressor in (2.3), as required by (2.2). The AR(1) in (2.3) can therefore be written on the same form as (2.1) and (2.2) with $\mathbf{x}_{i,t} = y_{i,t-1}$, $\mathbf{F}_t = (f_t, \alpha(L)f_{t-1})'$, $\boldsymbol{\lambda}_i = (\gamma_i, 0)'$, $\boldsymbol{\Lambda}_i = (0, \gamma_i)$, $\mathbf{v}_{i,t} = \alpha(L)\varepsilon_{i,t-1}$ and $\boldsymbol{\beta} = \alpha$. Hence, while the equation for $y_{i,t}$ in (2.3) has only one factor, the DGP for $\mathbf{y}_i$ and $\mathbf{x}_i$ in (2.1) and (2.2), respectively, has $r = 2$ factors.

Denote by $\mathbf{z}_i = (\mathbf{y}_i, \mathbf{x}_i) = (\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,T})'$ the $T \times (k + 1)$ matrix of observables. The DGP in (2.1) and (2.2) can be rewritten as the following static factor model for $\mathbf{z}_i$:

$$\mathbf{z}_i = \mathbf{F}\mathbf{C}_i + \mathbf{u}_i, \tag{2.5}$$

where $\mathbf{C}_i = (\boldsymbol{\Lambda}_i'\boldsymbol{\beta} + \boldsymbol{\lambda}_i, \boldsymbol{\Lambda}_i')$ is $r \times (k + 1)$ and $\mathbf{u}_i = (\mathbf{v}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \mathbf{v}_i) = (\mathbf{u}_{i,1}, \dots, \mathbf{u}_{i,T})'$ is $T \times (k + 1)$. This model is the same as the one considered by Pesaran (2006). The estimator of the factors is also the same, and is given by $\widehat{\mathbf{F}} = \bar{\mathbf{z}}$, where $\bar{\mathbf{A}} = N^{-1}\sum_{i=1}^{N}\mathbf{A}_i$ for any matrix $\mathbf{A}_i$. The CCEP estimator is simply the pooled OLS estimator with $\widehat{\mathbf{F}}$ in place of $\mathbf{F}$, which we may write as

$$\widehat{\boldsymbol{\beta}}_P = \left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{F}}}\mathbf{x}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{F}}}\mathbf{y}_i,$$

where $\mathbf{M}_{\mathbf{A}} = \mathbf{I}_T - \mathbf{A}(\mathbf{A}'\mathbf{A})^{+}\mathbf{A}' = \mathbf{I}_T - \mathbf{P}_{\mathbf{A}}$ for any T-rowed matrix $\mathbf{A}$ with $(\mathbf{A}'\mathbf{A})^{+}$ being the Moore–Penrose inverse of $\mathbf{A}'\mathbf{A}$ (if $\mathbf{A}$ is not of full column rank).

[2] Unlike in Pesaran (2006), the DGP considered in the present paper does not contain any observed factors, which are irrelevant for the main point. Observed factors can be accommodated simply by redefining $\mathbf{y}_i$, $\mathbf{x}_i$, $\mathbf{F}$ and $\boldsymbol{\varepsilon}_i$ as the residuals obtained by projecting on the observed factor matrix.

The assumptions that we will be working under are based on those of Pesaran (2006). The idea is to take the original assumptions of this seminal paper as a starting point, but to make a number of relaxations that will put stress on the CCEP approach. Before stating our first assumption, Assumption 2.1, it is useful to define

$$\mathbf{D} = \begin{pmatrix} 1 & \mathbf{0}'_{k\times 1} \\ \boldsymbol{\beta} & \mathbf{I}_k \end{pmatrix}$$

and $\mathbf{e}_{i,t} = (\varepsilon_{i,t}, \mathbf{v}'_{i,t})'$, which means that the errors in (2.5) can be written very conveniently as $\mathbf{u}_{i,t} = \mathbf{D}'\mathbf{e}_{i,t}$. Assumption 2.1 is stated in terms of the autocovariances of $\mathbf{e}_{i,t}$, which we are going to denote as

$$\boldsymbol{\Gamma}_{\mathbf{e},i}(h) = E(\mathbf{e}_{i,t}\mathbf{e}'_{i,t-h}) = \begin{pmatrix} E(\varepsilon_{i,t}\varepsilon_{i,t-h}) & E(\varepsilon_{i,t}\mathbf{v}'_{i,t-h}) \\ E(\mathbf{v}_{i,t}\varepsilon_{i,t-h}) & E(\mathbf{v}_{i,t}\mathbf{v}'_{i,t-h}) \end{pmatrix} = \begin{pmatrix} \Gamma_{\varepsilon,i}(h) & \boldsymbol{\Gamma}_{\varepsilon\mathbf{v},i}(h) \\ \boldsymbol{\Gamma}_{\varepsilon\mathbf{v},i}(-h)' & \boldsymbol{\Gamma}_{\mathbf{v},i}(h) \end{pmatrix}.$$

Assumption 2.1 (*Errors*)**.**

(i) $\mathbf{e}_{i,t}$ is a stationary process that is independent across i with $\boldsymbol{\Gamma}_{\mathbf{e},i}(h)$ absolutely summable, $E(\mathbf{e}_{i,t}) = \mathbf{0}_{(k+1)\times 1}$, $E(\|\mathbf{e}_{i,t}\|^4) < \infty$,

$$\boldsymbol{\Sigma}_{\mathbf{e},i} = \begin{pmatrix} \sigma^2_{\varepsilon,i} & \mathbf{0}'_{k\times 1} \\ \mathbf{0}_{k\times 1} & \boldsymbol{\Sigma}_{\mathbf{v},i} \end{pmatrix} = \boldsymbol{\Gamma}_{\mathbf{e},i}(0),$$

$\boldsymbol{\Sigma}_{\mathbf{e}} = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} \boldsymbol{\Sigma}_{\mathbf{e},i}$ is positive definite, and $\boldsymbol{\Gamma}_{\mathbf{e}}(h) = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} \boldsymbol{\Gamma}_{\mathbf{e},i}(h)$, where $\|\mathbf{A}\| = \sqrt{\mathrm{tr}\,(\mathbf{A}'\mathbf{A})}$ is the Frobenius norm of $\mathbf{A}$.

(ii) $\Gamma_{\varepsilon,i}(h) = 0$ for all $h \neq 0$.

(iii) $\boldsymbol{\Gamma}_{\varepsilon\mathbf{v},i}(h) = \mathbf{0}_{k\times 1}$ for all $h \geq 0$.

(iv) $\mathbf{v}_{i,t}$ and $\varepsilon_{j,s}$ are independent for all t, s and $i \neq j$.

Assumption 2.2 (*Factors*)**.**

(i) $\mathbf{F}_t$ is a stationary process with $E(\mathbf{F}_t) = \mathbf{0}_{r\times 1}$, $\boldsymbol{\Gamma}_{\mathbf{F}}(h) = \mathrm{plim}_{T\to\infty}\, T^{-1}\sum_{t=h+1}^{T} \mathbf{F}_t\mathbf{F}'_{t-h}$, $\boldsymbol{\Sigma}_{\mathbf{F}} = \boldsymbol{\Gamma}_{\mathbf{F}}(0)$ positive definite, and $\boldsymbol{\Gamma}_{\mathbf{F}}(h)$ absolutely summable.

(ii) $\mathbf{e}_{i,t}$ and $\mathbf{F}_s$ are mutually independent for all i, t and s.

The first thing to note about Assumption 2.1 is that it does not require $\boldsymbol{\Gamma}_{\varepsilon\mathbf{v},i}(h)$ to be zero for $h \leq 0$ (condition (iii)), which means that $\mathbf{x}_{i,t}$ may be predetermined such that $E(\varepsilon_{i,t}\mathbf{x}_{i,s}) = \mathbf{0}_{k\times 1}$ for $s \leq t$ and $E(\varepsilon_{i,t}\mathbf{x}_{i,s}) \neq \mathbf{0}_{k\times 1}$ for $s > t$. Hence, unlike in most other CCE studies (see, for example, Chudik et al., 2011; Karabiyik et al., 2017; Pesaran, 2006; Westerlund and Urbain, 2015, and Westerlund, 2018), here exogeneity fails not just because of the dependence among the common components of (2.1) and (2.2), but also because the idiosyncratic errors may be correlated with future values of the regressors. Chudik and Pesaran (2015), and Everaert and De Groote (2016) do allow for a lagged dependent variable, and are therefore more general in this regard. Assumption 2.1(iii) is, however, even more general, as it does not impose any restrictions on the nature of the predeterminedness, and is as a result applicable to a wider class of models that generate correlation between idiosyncratic errors and future values of the regressors (see Moon and Weidner, 2017 for a similar condition in case of QML estimation). For this to be possible, unlike in Pesaran (2006), we have to assume that $\varepsilon_{i,t}$ is serially uncorrelated (Assumption 2.1(ii)). Serial correlation can be permitted but then $\mathbf{x}_{i,t}$ can no longer be predetermined. Assumption 2.2(i) rules out non-stationary factors, which is a standard requirement in the literature (see Kapetanios et al., 2011, and Westerlund, 2018 for two exceptions). The requirement that the factors have zero mean is without loss of generality, as we can always (time) demean the data to ensure zero mean factors. The only difference is that in this case Assumption 2.2 should be appropriately reformulated in terms of $\mathbf{F}_t - T^{-1}\sum_{s=1}^{T} \mathbf{F}_s$, and not in terms of $\mathbf{F}_t$.

Assumption 2.3 (*Asymptotics*)**.** $N/T \to \kappa$ as $N, T \to \infty$ jointly with $\kappa \in (0, \infty)$.

Assumption 2.3 bounds the relative expansion rate of N and T such that $\kappa \neq 0$ and $\kappa^{-1} \neq 0$. In the QML strand of the literature, this is a common requirement (see, for example, Bai, 2009; Moon and Weidner, 2015, 2017, and Fernández-Val and Weidner, 2016). Pesaran (2006) considers both equal and heterogeneous slopes. In the former case, which is the one considered here, he assumes that $\kappa^{-1} = 0$. However, as the Monte Carlo results of Westerlund and Urbain (2015) clearly show, estimators constructed under this assumption are likely to suffer from poor small-sample properties. In this paper, we assume that $\kappa \neq 0$ and $\kappa^{-1} \neq 0$, which is only natural, because in many applications N and T are of similar magnitude.

As alluded to in Section 1, we would like to allow for the possibility that some of the factors that enter the equation for $\mathbf{x}_i$ do not enter the equation for $\mathbf{y}_i$, and vice verse. Assumption 2.4 allow $r - r_y$ of the factors in $\mathbf{F}$ not to enter the equation for $\mathbf{y}_i$. Using $\mathbf{F}_{-y}$ to denote these excluded factors, we may without loss of generality decompose $\mathbf{F} = (\mathbf{F}_y, \mathbf{F}_{-y})$, where $\mathbf{F}_y$ and $\mathbf{F}_{-y}$ are $T \times r_y$ and $T \times (r - r_y)$, respectively. The loading matrix $\mathbf{C}_i$ is partitioned conformably as $\mathbf{C}_i = (\mathbf{C}'_{y,i}, \mathbf{C}'_{-y,i})'$, where $\mathbf{C}_{y,i}$ is $r_y \times (k+1)$ and $\mathbf{C}_{-y,i}$ is $(r - r_y) \times (k+1)$. The loadings in (2.1) and (2.2) are partitioned analogously as $\boldsymbol{\lambda}_i = (\boldsymbol{\lambda}'_{y,i}, \boldsymbol{\lambda}'_{-y,i})'$ and $\boldsymbol{\Lambda}_i = (\boldsymbol{\Lambda}_{y,i}, \boldsymbol{\Lambda}_{-y,i})$, respectively. Note that while the factors in $\mathbf{F}_{-y}$ can only appear in the equation for $\mathbf{x}_i$, the factors in $\mathbf{F}_y$ can potentially appear in both equations. However, since the excluded factors are all in $\mathbf{F}_{-y}$, we

know that the factors in the equation for $\mathbf{y}_i$ are given by $\mathbf{F}_y$. These are the y-factors mentioned in Section 1. Assumption 2.4 formalizes this idea. The assumption is stated in terms of a full rank $(k+1)\times(k+1)$ matrix $\mathbf{Q}=(\mathbf{Q}_y, \mathbf{Q}_{-y})$, where $\mathbf{Q}_y$ and $\mathbf{Q}_{-y}$ are $(k+1)\times r_y$ and $(k+1)\times(k+1-r_y)$, respectively.

Assumption 2.4 (*Exclusion Restrictions*)**.**

(i) $\text{rk}(\bar{\mathbf{C}}_y\mathbf{Q}_y)=r_y$ for all N, including $N\to\infty$.
(ii) $\bar{\mathbf{C}}_{-y}\mathbf{Q}_y \overset{a.s.}{=} \mathbf{0}_{(r-r_y)\times r_y}$.
(iii) $\boldsymbol{\lambda}_{-y,i} \overset{a.s.}{=} \mathbf{0}_{(r-r_y)\times 1}$.

The significance of $\mathbf{Q}_y$ is that it enables us to rotate the columns of $\widehat{\mathbf{F}}$ into r_y columns that estimate $\mathbf{F}_y$ itself and $k+1-r_y$ columns that estimate a linear combination of both $\mathbf{F}_y$ and $\mathbf{F}_{-y}$. Note in particular how under Assumption 2.4(ii),

$$\begin{aligned}\widehat{\mathbf{F}}\mathbf{Q}_y &= \mathbf{F}\bar{\mathbf{C}}\mathbf{Q}_y + \bar{\mathbf{u}}\mathbf{Q}_y = (\mathbf{F}_y\bar{\mathbf{C}}_y + \mathbf{F}_{-y}\bar{\mathbf{C}}_{-y})\mathbf{Q}_y + \bar{\mathbf{u}}\mathbf{Q}_y \overset{a.s.}{=} \mathbf{F}_y\bar{\mathbf{C}}_y\mathbf{Q}_y + \bar{\mathbf{u}}\mathbf{Q}_y \\ &= \mathbf{F}_y\bar{\mathbf{C}}_y\mathbf{Q}_y + o_p(1). \end{aligned} \tag{2.6}$$

In fact, since $\bar{\mathbf{C}}_y\mathbf{Q}_y$ is invertible under Assumption 2.4(i), we have

$$\widehat{\mathbf{F}}\mathbf{Q}_y(\bar{\mathbf{C}}_y\mathbf{Q}_y)^{-1} \xrightarrow{p} \mathbf{F}_y \tag{2.7}$$

as $N\to\infty$ with T held fixed. In other words, $\widehat{\mathbf{F}}\mathbf{Q}_y(\bar{\mathbf{C}}_y\mathbf{Q}_y)^{-1}$ is consistent for $\mathbf{F}_y$. Of course, $\widehat{\mathbf{F}}\mathbf{Q}_y(\bar{\mathbf{C}}_y\mathbf{Q}_y)^{-1}$ is not really observable. However, since $\mathbf{M}_{\widehat{\mathbf{F}}}=\mathbf{M}_{\widehat{\mathbf{F}}\mathbf{Q}}$, for our purposes observing $\widehat{\mathbf{F}}$ is as good as observing $\widehat{\mathbf{F}}\mathbf{Q}_y(\bar{\mathbf{C}}_y\mathbf{Q}_y)^{-1}$. Assumption 2.4(i) and (ii) therefore ensure that $\widehat{\mathbf{F}}$ is useful for estimating $\mathbf{F}_y$. Assumption 2.4(iii) is needed to ensure that the y-factors are those that actually enters the equation for $\mathbf{y}_i$. The following example shows how Assumption 2.4 works in Example 2.1.

Example 2.1 (*Continued*)**.** In the pure AR(1) example, $\boldsymbol{\lambda}_i=(\gamma_i, 0)'$, and so $\boldsymbol{\lambda}_{-y,i}=0$, as required by Assumption 2.4(iii). But we also have $\boldsymbol{\Lambda}_i=(0,\gamma_i)$ and $\boldsymbol{\beta}=\alpha$, which means that $\mathbf{C}_i$ is given by

$$\mathbf{C}_i=(\boldsymbol{\Lambda}_i'\boldsymbol{\beta}+\boldsymbol{\lambda}_i, \boldsymbol{\Lambda}_i') = \begin{pmatrix}\gamma_i & 0\\ \alpha\gamma_i & \gamma_i\end{pmatrix}. \tag{2.8}$$

Moreover, since $r=k+1=2$ and $r_y=1$, we have $\mathbf{C}_{y,i}=(\gamma_i,0)$ and $\mathbf{C}_{-y,i}=(\alpha\gamma_i,\gamma_i)$. It follows that

$$\bar{\mathbf{C}}=\begin{pmatrix}\bar{\mathbf{C}}_y\\ \bar{\mathbf{C}}_{-y}\end{pmatrix}=\begin{pmatrix}\overline{\gamma} & 0\\ \alpha\overline{\gamma} & \overline{\gamma}\end{pmatrix}. \tag{2.9}$$

Consider $\mathbf{Q}$. Because $\mathbf{Q}$ is assumed to have full rank $k+1$, the $(k+1)\times r_y$ matrix $\mathbf{Q}_y$ must have full column rank r_y, which in this case means that the rank of $\mathbf{Q}_y$ should be one. We also require that $\text{rk}(\bar{\mathbf{C}}_y\mathbf{Q}_y)=r_y$ and $\bar{\mathbf{C}}_{-y}\mathbf{Q}_y \overset{a.s.}{=} \mathbf{0}_{(r-r_y)\times r_y}$. A natural choice is to set

$$\mathbf{Q}=(\mathbf{Q}_y,\mathbf{Q}_{-y})=\mathbf{D}^{-1}=\begin{pmatrix}1 & 0\\ -\alpha & 1\end{pmatrix}, \tag{2.10}$$

whose rank is obviously full with $\mathbf{Q}_y=(1,-\alpha)'$ having rank one. Also, $\bar{\mathbf{C}}_y\mathbf{Q}_y=\overline{\gamma}$ and $\bar{\mathbf{C}}_{-y}\mathbf{Q}_y=0$. Hence, if we assume that $\overline{\gamma}\neq 0$, then $\text{rk}(\bar{\mathbf{C}}_y\mathbf{Q}_y)=\text{rk}\,\overline{\gamma}=r_y=1$, and so Assumption 2.4 is met. The main restrictions here are therefore that $\boldsymbol{\lambda}_{-y,i}=0$ and $\overline{\gamma}\neq 0$ (for all N, including $N\to\infty$). The latter restriction makes sense, because if $\overline{\gamma}=0$, then $\bar{y}_t$ would not load on f_t, which means that its usefulness as an estimator would be lost, as we will explain in more detail later in this section.

As Example 2.1 makes clear, models in which the factors affecting the equations for $\mathbf{y}_i$ and $\mathbf{x}_i$ are different arise naturally in the presence of lagged dependent variables. It is therefore important to allow for this possibility.

Remark 2.1 (*The Factors in QML*)**.** The scenario that we are considering with both y-factors and non-y-factors can be motivated by the QML literature (see, for example, Bai, 2009 and Moon and Weidner, 2015, 2017). Here it is standard to assume that while $\mathbf{y}_i$ admits to a factor structure, the common component of $\mathbf{x}_i$ is essentially unrestricted. The idea is then to use QML to estimate and control for the factors in $\mathbf{y}_i$ only, which is again similar to our proposal. The main difference is that because of the way that the factors are estimated using the cross-sectional averages, in the current paper the common component of the regressors must have a factor structure.

Assumption 2.5 (*Loadings*)**.**

(i) $\mathbf{C}_i$ is independent across i with $E(\|\mathbf{C}_i\|^2)<\infty$ and $E(\mathbf{C}_i)=\mathbf{C}=(\mathbf{C}_y', \mathbf{C}_{-y}')'=(\boldsymbol{\Lambda}'\boldsymbol{\beta}+\boldsymbol{\lambda}, \boldsymbol{\Lambda}')$.
(ii) $\mathbf{C}_i$ and $\mathbf{u}_j$ are independent for all i and j.

(iii) $\mathbf{C} = \mathbf{R}\mathbf{C}_m$, where

$$\mathbf{R} = \begin{pmatrix} \mathbf{I}_{r_y} & \mathbf{0}_{r_y \times (r_m - r_y)} \\ \mathbf{0}_{(r-r_y)\times r_y} & \mathbf{R}_{22} \end{pmatrix}$$

and $\mathbf{C}_m$ are $r \times r_m$ and $r_m \times (k+1)$, respectively, with $\operatorname{rk}\mathbf{R} = \operatorname{rk}\mathbf{C}_m = r_m \geq r_y$.

Assumption 2.5 implies that

$$r_y \leq \operatorname{rk}\mathbf{C} = r_m \leq \min\{r, k+1\}. \tag{2.11}$$

As we will now explain, this condition is very general and can be seen as an extension of the rank condition first introduced by Pesaran (2006). The latter condition is by now standard in CCE studies with equal slopes (see Karabiyik et al., 2017 for an overview), and is given by

$$\operatorname{rk}\mathbf{C} = r = r_m = r_y \leq k+1. \tag{2.12}$$

To appreciate the difference between the two conditions it is useful to introduce the $T \times r_m$ matrix $\mathbf{F}_m = \mathbf{F}\mathbf{R}$, and the $T \times r$ matrix $\mathbf{F}_{-m} = \mathbf{M}_{\mathbf{F}_m}\mathbf{F}$. Note that $\mathbf{F}_m'\mathbf{F}_{-m} = \mathbf{0}_{r_m \times r}$, which means that the two sets of factors are orthogonal. The columns of $\mathbf{F}_m$ are the mean factors discussed in Section 1, while those of $\mathbf{F}_{-m}$ are the non-mean factors. Moreover, by making use of Assumption 2.5, the mean factors can be decomposed as $\mathbf{F}_m = (\mathbf{F}_y, \mathbf{F}_{m,-y})$, where $\mathbf{F}_{m,-y} = \mathbf{F}_{-y}\mathbf{R}_{22}$ is $T \times (r_m - r_y)$. Hence, the mean factors contain all of the y-factors in $\mathbf{F}_y$ plus $r_m - r_y$ linear combinations of the remaining $r - r_y$ non-y-factors in $\mathbf{F}_{-y}$. Note also that since $\mathbf{F}_y$ is included in $\mathbf{F}_m$, $\mathbf{F}_{-m}$ must have at least r_y zero columns. In fact, it is not difficult to show that $\mathbf{F}_{-m} = (\mathbf{0}_{T\times r_y}, \mathbf{F}_{-m,-y})$, where $\mathbf{F}_{-m,-y} = \mathbf{M}_{\mathbf{F}_m}\mathbf{F}_{-y}$ is $T \times (r - r_y)$.[3] Note also that if $r_m = r$, then $\mathbf{R}$ has full rank and so $\mathbf{M}_{\mathbf{F}_m} = \mathbf{M}_{\mathbf{F}}$, which in turn implies $\mathbf{F}_{-m,-y} = \mathbf{0}_{T\times(r-r_y)}$. This is to be expected, because if all the factors are mean factors, then by definition there cannot be any non-mean factors.

The definitions of $\mathbf{F}_m$ and $\mathbf{F}_{-m}$ imply that $\mathbf{F}$ has the following orthogonal components representation:

$$\mathbf{F} = \mathbf{F}_m\mathbf{H}' + \mathbf{F}_{-m}, \tag{2.13}$$

where $\mathbf{H} = \mathbf{F}'\mathbf{F}_m(\mathbf{F}_m'\mathbf{F}_m)^{-1}$ is $r \times r_m$. Clearly, $\mathbf{H}'\mathbf{R} = \mathbf{I}_{r_m}$, which in turn implies that $\mathbf{F}_{-m}\mathbf{R} = \mathbf{0}_{T\times r_m}$. By using this and Assumption 2.5,

$$\mathbf{F}\mathbf{C} = \mathbf{F}_m\mathbf{H}'\mathbf{R}\mathbf{C}_m + \mathbf{F}_{-m}\mathbf{R}\mathbf{C}_m = \mathbf{F}_m\mathbf{C}_m, \tag{2.14}$$

which in view of (2.5) in turn implies

$$E(\mathbf{z}_i|\mathbf{F}) = \mathbf{F}E(\mathbf{C}_i|\mathbf{F}) + E(\mathbf{u}_i|\mathbf{F}) = \mathbf{F}\mathbf{C} = \mathbf{F}_m\mathbf{C}_m. \tag{2.15}$$

Hence, in expectation there are not r but $r_m \leq r$ factors, which are identically the mean factors. As we show in (2.22), these are the factors that can be consistently estimated using the cross-sectional averages in $\widehat{\mathbf{F}}$. The remaining $r - r_m$ factors, which are the non-mean factors, are inestimable using the cross-sectional averages. This is intuitive, because factors that do not enter the expected value of $\mathbf{z}_i$ cannot reasonably be expected to be well-estimated by the sample average. By contrast, the condition in (2.12) requires that all factors are mean factors and that they can be consistently estimated. Consistent estimation of all factors is, however, not only very demanding but also unnecessary given our purpose to just estimate parameter β in the equation for $\mathbf{y}_i$. Condition (2.11) recognizes that there might be non-mean factors that only enter the equation for $\mathbf{x}_i$, and that these factors might not be estimable using cross-sectional averages.

Example 2.2 (*Mean and Non-mean Factors*)**.** The DGP that we will consider in this example is given by

$$y_{i,t} = \beta x_{i,t} + \gamma_i f_t + \varepsilon_{i,t}, \tag{2.16}$$

$$x_{i,t} = \pi_i f_t + \theta_i g_t + v_{i,t}. \tag{2.17}$$

This DGP is very similar to the production function considered by Eberhardt and Teal (2020), in which the inputs (labour and capital stock) may depend on factors that have no direct effect the value added. Another example of a DGP of this form is given by Cesa-Bianchi et al. (2019), where $y_{i,t}$ is the GDP growth and $x_{i,t}$ is a stock market volatility measure. Suppose that $E(\gamma_i) = \gamma$, $E(\pi_i) = \pi$ and $E(\theta_i) = \theta$, and let $\mathbf{f} = (f_1, \ldots, f_T)'$, $\mathbf{g} = (g_1, \ldots, g_T)'$ and $\mathbf{F} = (\mathbf{f}, \mathbf{g})$. In this notation, the y- and non-y-factors are given simply by $\mathbf{F}_y = \mathbf{f}$ and $\mathbf{F}_{-y} = \mathbf{g}$, respectively, with $\boldsymbol{\Lambda}_i = (\boldsymbol{\Lambda}_{y,i}, \boldsymbol{\Lambda}_{-y,i}) = (\pi_i, \theta_i)$ and $\boldsymbol{\lambda}_i = (\boldsymbol{\lambda}_{y,i}, \boldsymbol{\lambda}_{-y,i})' = (\gamma_i, 0)'$ being the associated loading matrices. It follows that

$$\mathbf{C} = (\boldsymbol{\Lambda}'\boldsymbol{\beta} + \boldsymbol{\lambda}, \boldsymbol{\Lambda}') = \begin{pmatrix} \pi\beta + \gamma & \pi \\ \theta\beta & \theta \end{pmatrix}, \tag{2.18}$$

[3] The fact that $\mathbf{F}_{-m} = (\mathbf{0}_{T\times r_y}, \mathbf{F}_{-m,-y})$ is a direct consequence of the block-wise formula for projection matrices, which states that $\mathbf{P}_{\mathbf{F}_m} = \mathbf{P}_{\mathbf{F}_y} + \mathbf{P}_{\mathbf{M}_{\mathbf{F}_y}\mathbf{F}_{m,-y}}$. It follows that $\mathbf{F}_{-m} = \mathbf{M}_{\mathbf{F}_m}\mathbf{F} = (\mathbf{M}_{\mathbf{F}_y} - \mathbf{P}_{\mathbf{M}_{\mathbf{F}_y}\mathbf{F}_{m,-y}})\mathbf{F} = (\mathbf{0}_{T\times r_y}, \mathbf{F}_{-m,-y})$, where $\mathbf{F}_{-m,-y} = \mathbf{M}_{\mathbf{F}_m}\mathbf{F}_{-y}$.

whose rank depends on whether or not $\theta = 0$. In this example, we set $\theta = 0$, which means that

$$\mathbf{C} = \begin{pmatrix} \pi\beta + \gamma & \pi \\ 0 & 0 \end{pmatrix}, \tag{2.19}$$

and therefore $r_y = r_m = 1 < r = k + 1 = 2$. One implication of this is that $\mathbf{R} = (1, 0)'$, which together with $\mathbf{C} = \mathbf{R}\mathbf{C}_m$ in Assumption 2.5(iii) in turn implies that $\mathbf{C}_m = (\pi\beta + \gamma, \pi)$. Hence, in this example, the mean and non-mean factors are given by $\mathbf{F}_m = \mathbf{F}\mathbf{R} = \mathbf{F}_y = \mathbf{f}$ and $\mathbf{F}_{-m} = \mathbf{M}_{\mathbf{f}}\mathbf{F} = (\mathbf{0}_{T\times 1}, \mathbf{M}_{\mathbf{f}}\mathbf{g}) = (\mathbf{0}_{T\times 1}, \mathbf{F}_{-m,-y})$, respectively.

Remark 2.2 (*Strong and Weak Factors*)**.** The concepts of mean and non-mean factors are related to the concepts of strong and weak factors introduced by Chudik et al. (2011). The main difference is that while the former concepts are about the mean of the loadings, the latter are about the order of the loadings themselves. In order to illustrate this difference, it is useful to consider the following very simple DGP:

$$y_{i,t} = \gamma_i f_t + \varepsilon_{i,t}, \tag{2.20}$$

where γ_i is independently distributed with mean $E(\gamma_i) = \gamma$. According to Definition 3.1 of Chudik et al. (2011), the factor f_t is strong if $\text{plim}_{N\to\infty} N^{-1}\sum_{i=1}^{N}|\gamma_i| > 0$, and it is weak if $\text{plim}_{N\to\infty}\sum_{i=1}^{N}|\gamma_i| < \infty$. The first condition will be satisfied if f_t is a mean factor, because then $\overline{\gamma} \xrightarrow{p} \gamma \neq 0$. If, on the other hand, f_t is a non-mean factor, then $\overline{\gamma} = o_p(1)$, which does not necessarily imply that $\sum_{i=1}^{N}|\gamma_i| = O_p(1)$. If, however, the factor is weak such that $\gamma_i = O_p(N^{-1})$, then it is also asymptotically a non-mean factor.

As the above discussion makes clear, one of the advantages of (2.11) is that not all factors have to be estimable. Another advantage is that r can be larger than $k + 1$, provided that $r_m \leq k + 1$. That is, the number of estimated factors may be smaller than the true number, which seems like a relevant scenario to consider, because in practice we never know how many factors there are. Moreover, while the regressors are typically given by economic theory, in many applications theory is not very informative about the factors (see, for example, Eberhardt et al., 2013). Therefore, the theoretically implied value of $k+1$ has typically little or nothing to do with r, and there is no reason to believe that the latter number should be less than or equal to the former. The fact (2.11) does not require that $r \leq k+1$ is therefore very valuable. The case when $r > k+1$ has been considered by a number of papers, including Chudik et al. (2011), Chudik and Pesaran (2015), Kapetanios et al. (2011), and Pesaran (2006). In these papers, however, the loadings are independent of the regressors, which, as already pointed out, means that the factors can be ignored. The number of factors is therefore irrelevant. There are some studies that consider DGPs with equal slopes; however, they all require that (2.12) holds (see Everaert and De Groote, 2016; Karabiyik et al., 2017; Pesaran, 2006; Westerlund and Urbain, 2015, and Westerlund, 2018). As far as we are aware, the current paper is the only one to consider a DGP with equal slopes and possibly correlated loadings without at the same time requiring that (2.12) holds.

2.2. Specifications

One of the challenges that we face while working under (2.11) is that r may be larger than r_m. This requires accounting for the non-mean factors in $\mathbf{F}_{-m}$, which are not captured by $\widehat{\mathbf{F}}$. Note in particular how

$$\widehat{\mathbf{F}} = \mathbf{F}\overline{\mathbf{C}} + \overline{\mathbf{u}} = \mathbf{F}_m\mathbf{H}'\overline{\mathbf{C}} + \mathbf{F}_{-m}(\overline{\mathbf{C}} - \mathbf{C}) + \overline{\mathbf{u}} = \mathbf{F}_m\overline{\mathbf{C}}_m + \overline{\mathbf{E}}, \tag{2.21}$$

where $\overline{\mathbf{C}}_m = \mathbf{H}'\overline{\mathbf{C}}$, $\overline{\mathbf{E}} = \mathbf{F}_{-m,-y}\widetilde{\mathbf{C}}_{-y} + \overline{\mathbf{u}}$ and $\widetilde{\mathbf{C}}_{-y} = \overline{\mathbf{C}}_{-y} - \mathbf{C}_{-y}$. The last equality makes use of the fact that $\mathbf{F}_{-m} = (\mathbf{0}_{T\times r_y}, \mathbf{F}_{-m,-y})$. The presence of $\mathbf{F}_{-m,-y}\widetilde{\mathbf{C}}_{-y}$ in $\overline{\mathbf{E}}$ is reflected in our results, which generally depend on $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$. That is, the results will in general depend on the unaccounted for non-mean factors.

While seemingly quite unproblematic when compared to the case when $r > r_m$, another major challenge we face while working under (2.11) is the fact that r_m may be smaller than $k+1$. In order to understand the issues involved, it is useful to first consider the case when $r_m = k + 1$, in which $\overline{\mathbf{C}}_m$ is $r_m \times r_m$ and invertible under Assumption 2.5(iii). Applying this to (2.21), we obtain

$$\widehat{\mathbf{F}}\,\overline{\mathbf{C}}_m^{-1} = \mathbf{F}_m + \overline{\mathbf{E}}\overline{\mathbf{C}}_m^{-1} \xrightarrow{p} \mathbf{F}_m. \tag{2.22}$$

Hence, not only are we able to estimate the y-factors consistently, but we are in fact able to estimate all the mean factors. The problem is if $r_m < k + 1$, which means that there are $k + 1 - r_m$ columns of $\widehat{\mathbf{F}}$ that do not load on $\mathbf{F}_m$. Hence, since $\overline{\mathbf{E}} \xrightarrow{p} \mathbf{0}_{T\times(k+1)}$ as $N \to \infty$, some of the columns of $\widehat{\mathbf{F}}$ will be degenerate, which in turn means that $T^{-1}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}$ appearing in the denominator of $\mathbf{M}_{\widehat{\mathbf{F}}}$ is singular in the limit. In order to side-step this problem, we look for a rotated version of $\widehat{\mathbf{F}}$ whose columns are all non-degenerate. Let us therefore decompose $\mathbf{Q} = (\mathbf{Q}_m, \mathbf{Q}_{-m})$, where $\mathbf{Q}_m$ and $\mathbf{Q}_{-m}$ are $(k + 1) \times r_m$ and $(k+1)\times(k+1-r_m)$, respectively. Note that the first r_y columns of $\mathbf{Q}_m$ are given by $\mathbf{Q}_y$. Hence, the matrix $\mathbf{Q}$ conveniently enables us to separate out both the y- and non-y-factors, and the mean and non-mean factors. Post-multiplying $\overline{\mathbf{C}}_m$ by this matrix, we obtain $\overline{\mathbf{C}}_m\mathbf{Q} = (\overline{\mathbf{C}}_m\mathbf{Q}_m, \overline{\mathbf{C}}_m\mathbf{Q}_{-m})$. Under Assumption 2.5(iii), the rank of $\overline{\mathbf{C}}_m$ is equal to r_m. It is therefore without loss of generality to assume that $\text{rk}(\overline{\mathbf{C}}_m\mathbf{Q}_m) = r_m$. Now let

$$\mathbf{G} = \begin{pmatrix} (\overline{\mathbf{C}}_m\mathbf{Q}_m)^{-1} & -\sqrt{N}(\overline{\mathbf{C}}_m\mathbf{Q}_m)^{-1}\overline{\mathbf{C}}_m\mathbf{Q}_{-m} \\ \mathbf{0}_{(k+1-r_m)\times r_m} & \sqrt{N}\mathbf{I}_{k+1-r_m} \end{pmatrix} = (\mathbf{G}_m, \mathbf{G}_{-m}), \tag{2.23}$$

where $\mathbf{G}_m$ is $(k+1)\times r_m$ and $\mathbf{G}_{-m}$ is $(k+1)\times(k+1-r_m)$. By construction, $\bar{\mathbf{C}}_m\mathbf{QG} = (\mathbf{I}_{r_m}, \mathbf{0}_{r_m\times(k+1-r_m)})$, from which it follows that

$$\widehat{\mathbf{F}}\mathbf{QG} = \mathbf{F}_m\bar{\mathbf{C}}_m\mathbf{QG} + \bar{\mathbf{E}}\mathbf{QG} = (\mathbf{F}_m, \mathbf{0}_{T\times(k+1-r_m)}) + \bar{\mathbf{E}}\mathbf{QG}. \tag{2.24}$$

The rotation by $\mathbf{QG}$ therefore rearranges the columns of $\widehat{\mathbf{F}}$ such that the first r_m columns of $\widehat{\mathbf{F}}\mathbf{QG}$ are those that load on $\mathbf{F}_m$. However, this is not the only thing that $\mathbf{QG}$ does. In order to appreciate this, note how $\bar{\mathbf{E}}\mathbf{QG} = (\bar{\mathbf{E}}\mathbf{QG}_m, \bar{\mathbf{E}}\mathbf{QG}_{-m})$, where $\bar{\mathbf{E}}\mathbf{QG}_m = \bar{\mathbf{E}}\mathbf{Q}_m(\bar{\mathbf{C}}_m\mathbf{Q}_m)^{-1}$ and $\bar{\mathbf{E}}\mathbf{QG}_{-m} = \sqrt{N}\bar{\mathbf{E}}[\mathbf{Q}_{-m} - \mathbf{Q}_m(\bar{\mathbf{C}}_m\mathbf{Q}_m)^{-1}\bar{\mathbf{C}}_m\mathbf{Q}_{-m}]$. Hence, because of the scaling by $\sqrt{N}$ in $\bar{\mathbf{E}}\mathbf{QG}_{-m}$, we have

$$\widehat{\mathbf{F}}\mathbf{QG} = (\mathbf{F}_m, \bar{\mathbf{E}}\mathbf{QG}_{-m}) + o_p(1), \tag{2.25}$$

where all the columns of the first term on the right are non-degenerate. Therefore, unlike $T^{-1}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}$, $T^{-1}\mathbf{G}'\mathbf{Q}'\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\mathbf{QG}$ converges to a positive definite matrix, which means that the singularity problem can be avoided. There is a problem, however, in that $\widehat{\mathbf{F}}\mathbf{QG}$ is not only estimating $\mathbf{F}_m$ but also $\bar{\mathbf{E}}\mathbf{QG}_{-m}$, which will affect the asymptotic distribution in very much the same way as the unaccounted for non-mean factors when $r > r_m$.

As the above discussion makes clear, the results of this paper depend critically on $k+1$, r_m and r. In what follows, we distinguish between the following four specifications:

S1. $k+1 = r_m = r$;
S2. $k+1 > r_m = r$;
S3. $k+1 = r_m < r$;
S4. $k+1 > r_m < r$.

As alluded to in Section 1, there exists a substantial variation in the DGPs considered in the existing literature, which makes it difficult to make general statements as to the generality of S1–S4. Consider S1 and S2. Similar specifications have been considered by, for example, Pesaran (2006), Karabiyik et al. (2017), Westerlund and Urbain (2015) and Westerlund (2018). However, none of these papers allow for predetermined regressors. Everaert and De Groote (2016) do allow for a lagged dependent variable, and are therefore more general in this regard. But then their DGP is basically the same single strong factor AR(1) as in Example 2.1, which represents a rather limited consideration. Hence, if we focus on existing studies that allow the loadings to be correlated with the regressors, even our simplest S1 and S2 specifications are more general than those considered previously.

Our interest in S3 originates with the fact that in the previous literature it is standard not to distinguish between r_m and r, which means that all factors are assumed to be mean factors. One exception here is the study of Chudik et al. (2011), who allow some of the factors to be weak with negligible loadings, which means that asymptotically they are non-mean factors (see Remark 2.2). However, since the weak factors are assumed to be uncorrelated with $\mathbf{x}_{i,t}$, they can be omitted without consequence. By contrast, as pointed out in the discussion following (2.21), the non-mean factors considered here will in general affect the asymptotic distribution of the CCEP estimator.

S4 is the most challenging scenario that we consider. Here we put no restrictions on the relationship between r and $k+1$, provided that they are both larger than r_m.

3. Asymptotic analysis

3.1. Asymptotic expansion

In this section, we provide an asymptotic expansion of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$, which, as already pointed out, is going to be key in the sequel. Before stating this expansion, however, it is useful to first introduce some notation. Specifically, we define

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}(h) &= \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{t=h+1}^{T}\mathbf{F}_{-m,-y,t}\mathbf{F}'_{-m,-y,t-h},\\
\boldsymbol{\Sigma}_{\bar{\mathbf{E}}} &= \sqrt{N}\,\widetilde{\mathbf{C}}'_{-y}\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}\sqrt{N}\,\widetilde{\mathbf{C}}_{-y} + \mathbf{D}'\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{D},\\
\mathbf{h} &= \boldsymbol{\Sigma}_{\mathbf{F}}\mathbf{J}_{\mathbf{F}}\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1},\\
\mathbf{J}_{\mathbf{F}} &= \mathbf{R}(\mathbf{R}'\boldsymbol{\Sigma}_{\mathbf{F}}\mathbf{R})^{-1}\mathbf{R}',\\
\mathbf{J}_{\bar{\mathbf{E}}} &= \begin{cases}\mathbf{Q}\mathbf{G}_{-m}(\mathbf{G}'_{-m}\mathbf{Q}'\boldsymbol{\Sigma}_{\bar{\mathbf{E}}}\mathbf{Q}\mathbf{G}_{-m})^{-1}\mathbf{G}'_{-m}\mathbf{Q}' & \text{for } k+1>r_m\\ \mathbf{0}_{(k+1)\times(k+1)} & \text{for } k+1=r_m\end{cases},
\end{aligned}$$

with $\mathbf{F}'_{-m,-y,t}$ being the $t-$th row of $\mathbf{F}_{-m,-y}$ and $\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}(0) = \boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}$. Analogous to $\widetilde{\mathbf{C}}_{-y}$, we further define $\widetilde{\boldsymbol{\Lambda}}_{-y,i} = \boldsymbol{\Lambda}_{-y,i} - \boldsymbol{\Lambda}_{-y}$.

Theorem 3.1. *Under Assumptions 2.1–2.5,*

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{b}_0 + \sqrt{\kappa}\mathbf{b}_1 + \kappa^{-1/2}\mathbf{b}_2 + \mathbf{b}_3 + \sqrt{T}\mathbf{b}_4) + o_p(1),$$

where

$$
\begin{aligned}
\mathbf{b}_0 &\xrightarrow{d} N(\mathbf{0}_{k\times 1}, \boldsymbol{\Sigma}_0),\\
\mathbf{b}_1 &= -\sum_{h=1}^{\infty} \boldsymbol{\Gamma}_{\varepsilon\mathbf{v}}(-h)' \mathrm{tr}\left[\boldsymbol{\Gamma}_{\mathbf{F}}(h)\mathbf{J}_{\mathbf{F}}\right],\\
\mathbf{b}_2 &= -\frac{1}{N}\sum_{i=1}^{N}[(\mathbf{0}_{k\times 1}, \mathbf{I}_k)\boldsymbol{\Sigma}_{\mathbf{e},i}\mathbf{D} - \boldsymbol{\Lambda}_i\mathbf{h}(\mathbf{C}_m\mathbf{Q}_m)^{-1\prime}\mathbf{Q}_m'\boldsymbol{\Sigma}_{\overline{\mathbf{E}}}](\boldsymbol{\Sigma}_{\overline{\mathbf{E}}}^{-1} - \mathbf{J}_{\overline{\mathbf{E}}})\\
&\quad\times \boldsymbol{\Sigma}_{\overline{\mathbf{E}}}\mathbf{Q}_m(\mathbf{C}_m\mathbf{Q}_m)^{-1}\mathbf{h}'(\boldsymbol{\lambda}_{y,i}', \mathbf{0}_{(r-r_y)\times 1}')'\\
&\quad - \frac{1}{N}\sum_{i=1}^{N}[(\mathbf{0}_{k\times 1}, \mathbf{I}_k)\boldsymbol{\Sigma}_{\mathbf{e},i}\mathbf{D} - \boldsymbol{\Lambda}_i\mathbf{h}(\mathbf{C}_m\mathbf{Q}_m)^{-1\prime}\mathbf{Q}_m'\boldsymbol{\Sigma}_{\overline{\mathbf{E}}}]\mathbf{J}_{\overline{\mathbf{E}}}(\sigma_{\varepsilon,i}^2, \mathbf{0}_{k\times 1}')'\\
&\quad - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Lambda}_i\mathbf{h}(\mathbf{C}_m\mathbf{Q}_m)^{-1\prime}\mathbf{Q}_m'(\sigma_{\varepsilon,i}^2, \mathbf{0}_{k\times 1}')',\\
\mathbf{b}_3 &\xrightarrow{d} N(\mathbf{0}_{k\times 1}, \boldsymbol{\Sigma}_3),\\
\mathbf{b}_4 &= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{\lambda}_{y,i}'\otimes\widetilde{\boldsymbol{\Lambda}}_{-y,i})\,\mathrm{vec}\left[\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{J}_{\overline{\mathbf{E}}}\mathbf{D}'\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{D}\mathbf{Q}_y(\mathbf{C}_y\mathbf{Q}_y)^{-1}\right],\\
\boldsymbol{\Sigma}_{\mathbf{x}} &= \boldsymbol{\Sigma}_{\mathbf{v}} + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Lambda}_{-y,i}(\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}} - \boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{J}_{\overline{\mathbf{E}}}\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}'\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}})\boldsymbol{\Lambda}_{-y,i}',
\end{aligned}
$$

which are all $O_p(1)$ *with* $\mathbf{b}_0$ *and* $\mathbf{b}_3$ *independent, having covariance matrices*

$$
\begin{aligned}
\boldsymbol{\Sigma}_0 &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sigma_{\varepsilon,i}^2\boldsymbol{\Sigma}_{\mathbf{v},i},\\
\boldsymbol{\Sigma}_3 &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sigma_{\varepsilon,i}^2 E(\widetilde{\boldsymbol{\Lambda}}_{-y,i}\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}\widetilde{\boldsymbol{\Lambda}}_{-y,i}')\\
&\quad + \left(\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}E[\boldsymbol{\lambda}_{y,i}'(\mathbf{C}_y\mathbf{Q}_y)^{-1\prime}\mathbf{Q}_y'\otimes\widetilde{\boldsymbol{\Lambda}}_{-y,i}]\right)\sum_{h=-\infty}^{\infty}[\mathbf{D}'\boldsymbol{\Gamma}_{\mathbf{e}}(h)\mathbf{D}\otimes\boldsymbol{\Gamma}_{\mathbf{F}_{-m,-y}}(h)]\\
&\quad\times\left(\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}E[\boldsymbol{\lambda}_{y,i}'(\mathbf{C}_y\mathbf{Q}_y)^{-1\prime}\mathbf{Q}_y'\otimes\widetilde{\boldsymbol{\Lambda}}_{-y,i}]\right)'.
\end{aligned}
$$

The expansion in Theorem 3.1 requires some discussion. Consider the numerator, which comprises five terms, $\mathbf{b}_0$, $\sqrt{\kappa}\mathbf{b}_1$, $\kappa^{-1/2}\mathbf{b}_2$, $\mathbf{b}_3$ and $\sqrt{T}\mathbf{b}_4$.[4] The first term is normal with mean zero, and will in many cases drive the asymptotic distribution of the estimator as a whole. The remaining terms are manifestations of the well-known "incidental parameters problem" (Neyman and Scott, 1948), which arises because of the need to account for the nuisance parameters in $(\mathbf{C}_1, \ldots, \mathbf{C}_N)$ and $(\mathbf{F}_1, \ldots, \mathbf{F}_T)$, whose number is increasing in N and T, respectively. Consider $\sqrt{\kappa}\mathbf{b}_1$. This is an asymptotically non-random term representing a bias in the asymptotic distribution of the numerator of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$. The source of this term is the bias coming from the estimation of $(\mathbf{C}_1, \ldots, \mathbf{C}_N)$ when T is fixed and the regressors are predetermined, which is a reflection of the "Nickell bias" (Nickell, 1981). The absolute magnitude of this bias depends on the specification considered. Note in particular that if $r = r_m$, as in S1 and S2, then $\mathbf{R}$ is $r_m \times r_m$ and invertible, which means that $\mathbf{J}_{\mathbf{F}} = \boldsymbol{\Sigma}_{\mathbf{F}}^{-1}$ and therefore $\mathbf{b}_1 = -\sum_{h=1}^{\infty}\boldsymbol{\Gamma}_{\varepsilon\mathbf{v}}(-h)'\mathrm{tr}\left[\boldsymbol{\Gamma}_{\mathbf{F}}(h)\boldsymbol{\Sigma}_{\mathbf{F}}^{-1}\right]$. If the regressors are exogenous, then $\boldsymbol{\Gamma}_{\varepsilon\mathbf{v}}(h)' = \mathbf{0}_{k\times 1}$, which means that $\mathbf{b}_1$ is zero, too. Similarly, if the factors are serially uncorrelated, then $\boldsymbol{\Gamma}_{\mathbf{F}}(h) = \mathbf{0}_{r\times r}$, and so $\mathbf{b}_1$ is again zero.

Because of the dependence on κ, the effect of $\sqrt{\kappa}\mathbf{b}_1$ is going to be less pronounced the larger is T relative to N. The opposite is true for $\kappa^{-1/2}\mathbf{b}_2$, which arises because of the need to account for $\mathbf{F}$. The fact that this term depends on $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$ through $\boldsymbol{\Sigma}_{\overline{\mathbf{E}}}$ (and hence also $\mathbf{J}_{\overline{\mathbf{E}}}$) means that it is generally random even in the limit as $N\to\infty$, and that it may therefore contribute to both the mean and variance of the asymptotic distribution of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$. An important consideration in this regard is whether $k+1 = r_m$ or $k+1 > r_m$. As alluded to in Section 2.2, the reason for this is the redundant factor estimates when $k+1 > r_m$, which affect the asymptotic behavior of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ in very much the same way as

[4] Consider the matrix product $A^{-1}B$ in which A and B are conformable matrices. In this paper, we refer to B as the "numerator" and A as the "denominator" even if their dimension is greater than one.

redundant deterministic constant and trend terms do in unit root testing. If, however, $k+1=r_m$ (as in S1 and S3), then $\mathbf{J}_{\bar{\mathbf{E}}}=\mathbf{0}_{(k+1)\times(k+1)}$, and so $\mathbf{b}_2$ reduces to

$$\begin{aligned}\mathbf{b}_2 &= -\frac{1}{N}\sum_{i=1}^{N}[(\mathbf{0}_{k\times 1},\mathbf{I}_k)\boldsymbol{\Sigma}_{\mathbf{e},i}\mathbf{D}-\boldsymbol{\Lambda}_i\mathbf{h}(\mathbf{C}_m\mathbf{Q}_m)^{-1\prime}\mathbf{Q}_m'\boldsymbol{\Sigma}_{\bar{\mathbf{E}}}]\\ &\times\ \mathbf{Q}_m(\mathbf{C}_m\mathbf{Q}_m)^{-1}\mathbf{h}'(\boldsymbol{\lambda}_{y,i}',\mathbf{0}_{(r-r_y)\times 1}')'-\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Lambda}_i\mathbf{h}(\mathbf{C}_m\mathbf{Q}_m)^{-1\prime}\mathbf{Q}_m'(\sigma_{\varepsilon,i}^2,\mathbf{0}_{k\times 1}')'\\ &= -\frac{1}{N}\sum_{i=1}^{N}[(\mathbf{0}_{k\times 1},\mathbf{I}_k)\boldsymbol{\Sigma}_{\mathbf{e},i}\mathbf{D}-\boldsymbol{\Lambda}_i\mathbf{h}\mathbf{C}_m^{-1\prime}\boldsymbol{\Sigma}_{\bar{\mathbf{E}}}]\mathbf{C}_m^{-1}\mathbf{h}'(\boldsymbol{\lambda}_{y,i}',\mathbf{0}_{(r-r_y)\times 1}')'\\ &-\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Lambda}_i\mathbf{h}\mathbf{C}_m^{-1\prime}(\sigma_{\varepsilon,i}^2,\mathbf{0}_{k\times 1}')',\end{aligned}\tag{3.1}$$

where the second equality holds, because $\mathbf{C}_m$ and $\mathbf{Q}_m$ are both $r_m\times r_m$ and invertible under $k+1=r_m$. Without further restrictions, this expression is nonzero.[5] Hence, $\mathbf{b}_2$ is there even in our most restrictive S1 specification, as to be expected, because of the estimated factors, which are always included. Note also how $\mathbf{b}_2$ depends on $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$ through $\boldsymbol{\Sigma}_{\bar{\mathbf{E}}}$. It is therefore random, as opposed to $\mathbf{b}_1$. However, there are cases when this randomness goes away. One is when $r=r_m$, as in S1 and S2, in which case $\mathbf{F}_{-m,-y}=\mathbf{0}_{T\times(r-r_y)}$ and therefore $\boldsymbol{\Sigma}_{\bar{\mathbf{E}}}=\mathbf{D}'\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{D}$. Another instance when $\mathbf{b}_2$ is non-random is in S3.

Proposition 3.1. *Suppose that the conditions of Theorem 3.1 hold. Then, under S3,*

$$\begin{aligned}\mathbf{b}_2 &= -\frac{1}{N}\sum_{i=1}^{N}[(\mathbf{0}_{k\times 1},\mathbf{I}_k)\boldsymbol{\Sigma}_{\mathbf{e},i}-\boldsymbol{\Lambda}_i\mathbf{h}\mathbf{C}_m^{-1\prime}\mathbf{D}'\boldsymbol{\Sigma}_{\mathbf{e}}]\mathbf{D}\mathbf{C}_m^{-1}\mathbf{h}'(\boldsymbol{\lambda}_{y,i}',\mathbf{0}_{(r-r_y)\times 1}')'\\ &-\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Lambda}_i\mathbf{h}\mathbf{C}_m^{-1\prime}(\sigma_{\varepsilon,i}^2,\mathbf{0}_{k\times 1}')'.\end{aligned}$$

Note that the expression given in Proposition 3.1 does not depend on $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$. It follows that while generally random in S4, in S1–S3 $\mathbf{b}_2$ is deterministic. This is going to be important in Section 3.3 where we study the asymptotic distribution of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P-\boldsymbol{\beta})$.

The last two terms in the numerator, $\mathbf{b}_3$ and $\sqrt{T}\mathbf{b}_4$, are zero in S1 and S2, as in these specifications $r=r_m$, and so $\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}=\mathbf{0}_{(r-r_y)\times(r-r_y)}$. The reason for the presence of these terms in S3 and S4 is the non-mean factors, which are not captured by $\widehat{\mathbf{F}}$ and that can be seen as omitted variables. In S3, these omitted variables are asymptotically uncorrelated with both $\mathbf{x}_i$ and $\widehat{\mathbf{F}}$. This implies that $\mathbf{J}_{\bar{\mathbf{E}}}=\mathbf{0}_{(k+1)\times(k+1)}$, and therefore $\mathbf{b}_4=\mathbf{0}_{k\times 1}$. In S4, however, the redundant factor estimates are correlated with the omitted factors, which causes the effect to increase and to become rather serious. In particular, since $\mathbf{b}_4=O_p(1)$, $\sqrt{T}\mathbf{b}_4$ is divergent and therefore so is $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P-\boldsymbol{\beta})$.

As with the numerator, the exact form of the denominator depends on the specification considered. Note in particular that while in S1 and S2 $\boldsymbol{\Sigma}_{\mathbf{x}}=\boldsymbol{\Sigma}_{\mathbf{v}}$, in S3 and S4 there are additional terms that depend on whether $k+1>r_m$ and/or $r>r_m$. Note in particular how $\boldsymbol{\Sigma}_{\mathbf{x}}$ generally depend on $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$.

Remark 3.1 (*Comparison with the QML Bias*)**.** Theorem 3.1 bears many similarities to the results obtained by Bai (2009), and Moon and Weidner (2015, 2017) for their QML estimator. Note in particular how $\mathbf{b}_1$, the Nickell bias, is present also in the asymptotic representation of this other estimator. However, since the QML estimator is only concerned with those factors that actually enter the equation for $\mathbf{y}_i$, the resulting bias only depends on $\mathbf{F}_y$. The comparison of the other terms is less straightforward (see Westerlund and Urbain, 2015). However, we note that unlike in CCE, which does not make use of the information regarding the covariance structure of $\boldsymbol{\varepsilon}_i$, the non-Nickell part of the bias of the QML estimator is zero in absence of cross-sectional and time series heteroscedasticity in $\boldsymbol{\varepsilon}_i$.

The asymptotic expansion of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P-\boldsymbol{\beta})$ simplifies considerably if a specific DGP can be assumed. This is illustrated in the following examples.

Example 2.1 (*Continued*)**.** Suppose that the AR(1) given earlier in this example holds, that $\sigma_{\varepsilon,i}^2=\sigma_{\varepsilon}^2$ for all i, and that $f_t=\phi f_{t-1}+\eta_t$ with η_t being mean zero and independent of $\varepsilon_{i,s}$ for all t, i and s. Under these assumptions, $\boldsymbol{\Gamma}_{\boldsymbol{\varepsilon}\mathbf{v}}(-h)'=\alpha^{h-1}\sigma_{\varepsilon}^2$ and $\mathbf{F}_t=(f_t,\alpha(L)f_{t-1})'$ has the following first-order vector autoregressive, or VAR(1), representation:

$$\mathbf{F}_t=\begin{pmatrix}\phi & 0\\ 1 & \alpha\end{pmatrix}\mathbf{F}_{t-1}+\begin{pmatrix}1\\ 0\end{pmatrix}\eta_t.\tag{3.2}$$

[5] Below we provide an example of a situation in which $\mathbf{b}_2=\mathbf{0}_{k\times 1}$.

The fact that $\mathbf{F}_t$ admits to a VAR(1) representation implies that

$$\mathbf{\Gamma}_{\mathbf{F}}(h) = \begin{pmatrix} \phi & 0 \\ 1 & \alpha \end{pmatrix}^h \mathbf{\Gamma}_{\mathbf{F}}(0), \tag{3.3}$$

from which it follows that $\mathrm{tr}\left[\mathbf{\Gamma}_{\mathbf{F}}(h)\mathbf{J}_{\mathbf{F}}\right] = \phi^h + \alpha^h$. Direct insertion into $\mathbf{b}_1$ now yields

$$\mathbf{b}_1 = -\sum_{h=1}^{\infty} \mathbf{\Gamma}_{\varepsilon\mathbf{v}}(-h)'\mathrm{tr}\left[\mathbf{\Gamma}_{\mathbf{F}}(h)\mathbf{J}_{\mathbf{F}}\right] = -\sigma_{\varepsilon}^2\left(\frac{\phi}{1-\alpha\phi} + \frac{\alpha}{1-\alpha^2}\right). \tag{3.4}$$

This differs from the bias of the infeasible OLS estimator obtained by taking f_t as known, which is given by $-\sigma_{\varepsilon}^2\phi(1-\alpha\phi)^{-1}$. Thus, since in many empirically relevant cases, both α and ϕ are expected to be positive, the bias of the CCEP estimator is going to be larger than that of the infeasible OLS estimator. The CCEP bias is also going to be larger than the corresponding QML bias, which is identical to the infeasible OLS bias, see Moon and Weidner (2017).

Example 2.1 (*Continued*)**.** Let us now drop the AR(1) specification for f_t and go back to the original AR(1) model for $y_{i,t}$. In this model, $r_y = 1$ and $r_m = r = k+1 = 2$. Making use of the fact that $r_m = r$, we obtain $\mathbf{\Sigma}_{\bar{\mathbf{E}}} = \mathbf{D}'\mathbf{\Sigma}_{\mathbf{e}}\mathbf{D}$. Also, $\mathbf{C} = (\boldsymbol{\lambda}, \mathbf{\Lambda}')\mathbf{D} = \gamma\mathbf{D}$. Insertion into (3.1) yields

$$\mathbf{b}_2 = -\frac{1}{N}\sum_{i=1}^{N}[(0,1)\mathbf{\Sigma}_{\mathbf{e},i} - \gamma^{-1}\mathbf{\Lambda}_i\mathbf{\Sigma}_{\mathbf{e}}]\gamma^{-1}(\boldsymbol{\lambda}_{y,i}, 0)' - \frac{1}{N}\sum_{i=1}^{N}\gamma^{-1}\mathbf{\Lambda}_i\mathbf{D}^{-1\prime}(\sigma_{\varepsilon,i}^2, 0)'. \tag{3.5}$$

We also have $\mathbf{\Lambda}_i = \gamma_i(0,1)$, $\boldsymbol{\lambda}_{y,i} = \gamma_i$, and

$$\mathbf{D}^{-1} = \begin{pmatrix} 1 & \mathbf{0} \\ -\alpha & 1 \end{pmatrix}, \tag{3.6}$$

where the latter implies $(0,1)\mathbf{D}^{-1\prime}(1,0)' = 0$. Similarly, because $\mathbf{\Sigma}_{\mathbf{e},i}$ and $\mathbf{\Sigma}_{\mathbf{e}}$ are both diagonal, we have $(0,1)\mathbf{\Sigma}_{\mathbf{e},i}(1,0)' = (0,1)\mathbf{\Sigma}_{\mathbf{e}}(1,0)' = 0$. It follows that

$$\mathbf{b}_2 = -\frac{1}{N}\sum_{i=1}^{N}[(0,1)\mathbf{\Sigma}_{\mathbf{e},i} - \gamma^{-1}\gamma_i(0,1)\mathbf{\Sigma}_{\mathbf{e}}]\gamma^{-1}(\gamma_i, 0)' - \frac{1}{N}\sum_{i=1}^{N}\gamma^{-1}\gamma_i(0,1)\mathbf{D}^{-1\prime}(\sigma_{\varepsilon,i}^2, 0)' = 0. \tag{3.7}$$

Moreover, since $\mathbf{b}_3$ and $\mathbf{b}_4$ are clearly zero, we have that in the particular AR(1) model considered here there is only the Nickell bias. One can similarly show that the same conclusion applies to VAR(p) models of the type considered by Chudik and Pesaran (2015).

3.2. Consistency

Theorem 3.1 has obvious implications for consistency. These are summarized in Corollaries 3.1 and 3.2. We begin by considering the results for S1–S3, which are given in Corollary 3.1. This corollary is a simple consequence of the fact that κ, κ^{-1} and $\mathbf{b}_0$–$\mathbf{b}_3$ are all $O_p(1)$, and that $\mathbf{b}_4 = \mathbf{0}_{k\times 1}$ in S1–S3.

Corollary 3.1. *Suppose that the conditions of Theorem 3.1 hold. Then, under S1–S3,*

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(1).$$

Corollary 3.1 states that the CCEP estimator is consistent and that the rate of convergence is given by $\sqrt{NT}$. This result is in line with previous results for DGPs with equal slopes (see, for example, Pesaran, 2006; Karabiyik et al., 2017; Westerlund and Urbain, 2015, and Westerlund, 2018), many of which were derived under more restrictive conditions than the ones considered here. Note in particular how these existing results assume that (2.12) holds, and that the regressors are exogenous. Corollary 3.1 is more general in this regard and can therefore be seen an extension of previous works.

As for S4, according to Theorem 3.1, we have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = \mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{b}_4 + O_p(T^{-1/2}). \tag{3.8}$$

The next corollary is a direct consequence of this.

Corollary 3.2. *Suppose that the conditions of Theorem 3.1 hold. Then, under S4,*

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(1).$$

According to Corollaries 3.1 and 3.2, while when at most one of $r > r_m$ and $k+1 > r_m$ are entertained (as in S1–S3) the rate of convergence is $\sqrt{NT}$, when both are entertained the rate is only $\sqrt{N}$. Note also that while Corollary 3.2 makes

use of Assumption 2.3, $\sqrt{N}/T \to 0$ is actually enough for $\sqrt{N}$-consistency under S4. Without restrictions on the relative expansion rate of N and T, we have the following result:

$$\sqrt{\min\{N,T\}}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(1). \tag{3.9}$$

The only condition here is therefore that $\min\{N,T\} \to \infty$.

While under the conditions of Theorem 3.1 the rate of convergence in S4 is $\sqrt{N}$, there are exceptions when the rate of convergence is $\sqrt{NT}$. In what follows, we will consider four such exceptions, which are given in Assumption 3.1(i)–(iv).

Assumption 3.1 (*Extra Loading Restrictions*)**.** One of the following conditions hold:

(i) $E(\widetilde{\boldsymbol{\Lambda}}_{-y,i}\boldsymbol{\lambda}_{y,i}) = \mathbf{0}_{k\times 1}$;
(ii) $\widetilde{\boldsymbol{\Lambda}}_{-y,i} = O_p(N^{-1/4})$ and $\widetilde{\boldsymbol{\Lambda}}_{-y} = O_p(N^{-3/4})$;
(iii) $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q} = O_p(N^{-1/2})$;
(iv) $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q}_{-m} = O_p(N^{-1/2})$ and $\overline{\mathbf{C}}_m\mathbf{Q}_{-m} = O_p(N^{-1/2})$.

According to Assumption 3.1(i), one exception from the $\sqrt{N}$-rate of convergence in S4 is when $\boldsymbol{\Lambda}_{-y,i}$ and $\boldsymbol{\lambda}_{y,i}$ are uncorrelated. This is intuitive, as the cause of the reduced convergence rate in S4 is the non-y-factors, which are not fully captured in the estimation. The y-factors are unobserved, but we know that they can be consistently estimated using $\widehat{\mathbf{F}}$ (or rather $\widehat{\mathbf{F}}\mathbf{Q}_y(\overline{\mathbf{C}}_y\mathbf{Q}_y)^{-1}$). Hence, asymptotically having $\widehat{\mathbf{F}}$ is just as good as having $\mathbf{F}_y$ itself. The factors included in the model are therefore correlated with those not captured in $\mathbf{x}_i$, leading to an "omitted variables bias", as represented by $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{b}_4$ in (3.8). By assuming that the loadings of the y-factors in (2.1) are uncorrelated with the non-y-factor loadings in (2.2), we obtain $N^{-1}\sum_{i=1}^{N}(\boldsymbol{\lambda}_{y,i}'\otimes\widetilde{\boldsymbol{\Lambda}}_{-y,i}) = O_p(N^{-1/2})$, implying that $\mathbf{b}_4$ is of the same order. Hence, in view of (3.8) and Assumption 2.3, it is easy to see that

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(\sqrt{T}N^{-1/2}) + O_p(1) = O_p(1). \tag{3.10}$$

In other words, by assuming that $\boldsymbol{\lambda}_{y,i}$ and $\boldsymbol{\Lambda}_{-y,i}$ are uncorrelated, the size of the bias is reduced and, as a result, $\sqrt{NT}$-consistency is restored. Westerlund and Urbain (2013) recognize the importance of having uncorrelated loadings when $k+1 < r$. However, they assume that the whole of $\boldsymbol{\Lambda}_i$ and $\boldsymbol{\lambda}_i$ are uncorrelated (as do Chudik et al., 2011; Kapetanios et al., 2011, and Pesaran, 2006), which in the current context is too restrictive. In fact, if $\boldsymbol{\Lambda}_i$ and $\boldsymbol{\lambda}_i$ are uncorrelated, even simple fixed effects OLS is consistent, and so there is really no need to use CCE.

Another exception from the relatively low rate of convergence in S4 is when the variation in $\boldsymbol{\Lambda}_{-y,i}$ is local-to-zero. Assumption 3.1(ii) requires that $\widetilde{\boldsymbol{\Lambda}}_{-y,i} = O_p(N^{-1/4})$ and $\widetilde{\boldsymbol{\Lambda}}_{-y} = N^{-1}\sum_{i=1}^{N}\widetilde{\boldsymbol{\Lambda}}_{-y,i} = O_p(N^{-3/4})$, which will be the case if $\widetilde{\boldsymbol{\Lambda}}_{-y,i} = N^{-1/4}\widetilde{\boldsymbol{\Lambda}}^*_{-y,i}$, where $\boldsymbol{\Lambda}^*_{-y,i}$ satisfies the conditions previously placed on $\boldsymbol{\Lambda}_{-y,i}$. This is similar to the weak factor case discussed in Remark 2.2, but is less restrictive, as $\boldsymbol{\Lambda}_{-y}$ does not have to be shrinking to zero. In order to appreciate the effect of this assumption, note that by Assumption 2.4(iii), $\widetilde{\mathbf{C}}_{-y} = (\widetilde{\boldsymbol{\Lambda}}_{-y}'\boldsymbol{\beta}+\widetilde{\boldsymbol{\lambda}}_{-y}, \widetilde{\boldsymbol{\Lambda}}_{-y}') \overset{a.s.}{=} (\widetilde{\boldsymbol{\Lambda}}_{-y}'\boldsymbol{\beta}, \widetilde{\boldsymbol{\Lambda}}_{-y}')$, which is clearly of the same order as $\widetilde{\boldsymbol{\Lambda}}_{-y}$. Hence, because of the way that $\mathbf{b}_4$ depends on both $\widetilde{\boldsymbol{\Lambda}}_{-y,i}$ and $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$, we have that $\mathbf{b}_4 = O_p(N^{-1/2})$, which means that $\sqrt{NT}$-consistency is again restored.

Assumption 3.1(i) and (ii) are useful because they give easy-to-interpret conditions under which $\sqrt{NT}$-consistency holds. The main drawback is that they apply directly to $\boldsymbol{\Lambda}_{-y,i}$ and $\boldsymbol{\lambda}_{y,i}$, which can be restrictive. Another possibility is to put restrictions not on the loadings themselves but on their rotated versions. As already pointed out, the reason for the low convergence rate in S4 is the presence of $\mathbf{b}_4$. The point that we would like to make here is that if we are not willing to restrict $\boldsymbol{\Lambda}_{-y,i}$ or $\boldsymbol{\lambda}_{y,i}$, then we have to restrict

$$\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{J}_{\overline{\mathbf{E}}} = \sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q}N^{-1/2}\mathbf{G}_{-m}(N^{-1/2}\mathbf{G}_{-m}'\mathbf{Q}'\boldsymbol{\Sigma}_{\overline{\mathbf{E}}}\mathbf{Q}N^{-1/2}\mathbf{G}_{-m})^{-1}N^{-1/2}\mathbf{G}_{-m}'\mathbf{Q}. \tag{3.11}$$

Since $(N^{-1/2}\mathbf{G}_{-m}'\mathbf{Q}'\boldsymbol{\Sigma}_{\overline{\mathbf{E}}}\mathbf{Q}N^{-1/2}\mathbf{G}_{-m})^{-1}$ and $N^{-1/2}\mathbf{G}_{-m}'\mathbf{Q}$ are both $O_p(1)$, we focus on

$$\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q}N^{-1/2}\mathbf{G}_{-m} = \sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q}_{-m} - \sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q}_m(\overline{\mathbf{C}}_m\mathbf{Q}_m)^{-1}\overline{\mathbf{C}}_m\mathbf{Q}_{-m}. \tag{3.12}$$

If this matrix is zero, then $\widetilde{\mathbf{C}}_{-y}\mathbf{J}_{\overline{\mathbf{E}}}$ and hence $\mathbf{b}_4$ will be zero, too. Of course, zero restrictions are quite strong, and so we instead restrict the probabilistic order of $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q}N^{-1/2}\mathbf{G}_{-m}$. Two possibilities arise.

One possibility is to assume that $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q} = O_p(N^{-1/2})$, as in Assumption 3.1(iii), which is a restriction on the variance of some of the loadings. Specifically, the assumption states that the variance of the non-y-factor loadings in $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$ is negligible after rotation by $\mathbf{Q}$. This makes sense, because it is exactly the unaccounted for non-y-factors in $\mathbf{x}_i$ that cause the problem. By assuming that the variance of the loadings goes to zero, the effect of these factors is reduced. Note that in contrast to the local-to-zero $\widetilde{\boldsymbol{\Lambda}}_{-y,i}$ scenario, here $\widetilde{\boldsymbol{\Lambda}}_{-y,i}$ is not necessarily negligible just because $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q}$ is, which explains the relatively high rate of shrinking in the current scenario.

Alternatively, we may assume that $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}\mathbf{Q}_{-m}$ and $\sqrt{N}\overline{\mathbf{C}}_m\mathbf{Q}_{-m}$ are $O_p(N^{-1/2})$, as in Assumption 3.1(iv). This imposes fewer variance restrictions, at the cost of additional restrictions on $\overline{\mathbf{C}}_m$. Specifically, it is assumed that those columns of $\mathbf{z}_i\mathbf{Q}$ that are not useful for estimating the mean factors in $\mathbf{F}_m$ have negligible loadings. The logic here is similar to before;

the y-factors in (2.1) are also mean factors, which are correlated with the unaccounted for non-mean factors in (2.2), and the negligible loading restriction reduces the effect of these non-mean factors.

Proposition 3.2 shows that the rate of convergence under Assumption 3.1 is indeed given by $\sqrt{NT}$.

Proposition 3.2. *Suppose that the conditions of Corollary 3.2 hold. If in addition Assumption 3.1 is met, then*

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(1).$$

Example 2.2 (*Continued*)**.** In this example, we illustrate the implications of Assumption 3.1 for the DGP considered in Example 2.2, where $r_y = r_m = 1 < r = k + 1 = 2$, which means that S4 holds. Our starting point is the following much simplified expression for $\mathbf{b}_4$:

$$\mathbf{b}_4 = -\frac{(N^{-1}\sum_{i=1}^{N}\theta_i\gamma_i)(\text{plim}_{T\to\infty}T^{-1}\mathbf{g}'\mathbf{M}_\mathbf{f}\mathbf{g})\sqrt{N}\overline{\theta}\overline{\pi}\sigma_\varepsilon^2}{\gamma[\overline{\gamma}^2\boldsymbol{\Sigma}_\mathbf{v} + \overline{\gamma}^2(\sqrt{N}\overline{\theta})^2(\text{plim}_{T\to\infty}T^{-1}\mathbf{g}'\mathbf{M}_\mathbf{f}\mathbf{g}) + \overline{\pi}^2\sigma_\varepsilon^2]}, \tag{3.13}$$

which has been obtained by direct substitution into the general formula given in Theorem 3.1. This expression is $O_p(1)$ without further assumptions, which in view of (3.8) means that $\sqrt{N}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ is of the same order. One possibility in order to bring the order of $\mathbf{b}_4$ down is to assume that $\boldsymbol{\Lambda}_{-y,i} = \theta_i$ and $\boldsymbol{\lambda}_{y,i} = \gamma_i$ are uncorrelated, as required by Assumption 3.1(i). This means that $N^{-1}\sum_{i=1}^{N}\theta_i\gamma_i = O_p(N^{-1/2})$, such that $\mathbf{b}_4 = O_p(N^{-1/2})$. Insertion into (3.8) yields $\sqrt{N}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(N^{-1/2}) + O_p(T^{-1/2})$, which under Assumption 2.3 is equivalent to $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(1)$. Another way to achieve the same goal is to assume that $\theta_i = O_p(N^{-1/4})$, such that θ_i is local-to-zero. This is Assumption 3.1(ii). For Assumption 3.1(iii), we note that $\widetilde{\mathbf{C}}_{-y}\mathbf{Q} = (\overline{\theta}\beta, \overline{\theta})$ and, analogously to Example 2.1,

$$\mathbf{Q} = (\mathbf{Q}_m, \mathbf{Q}_{-m}) = \mathbf{D}^{-1} = \begin{pmatrix} 1 & 0 \\ -\beta & 1 \end{pmatrix}, \tag{3.14}$$

implying that $\widetilde{\mathbf{C}}_{-y}\mathbf{Q} = (0, \overline{\theta})$. Hence, for condition (iii) to be met we require $\overline{\theta} = O_p(N^{-1})$, which again means that $\mathbf{b}_4 = O_p(N^{-1/2})$, and so $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(1)$. One way to satisfy (iii) is to set $\theta_i = N^{-1/2}\theta_i^*$ with θ_i^* mean zero and independent across i, which is similar to (ii) but in this particular example more restrictive, as the rate of shrinking is relatively higher. One of the requirements of Assumption 3.1(iv) is that $\widetilde{\mathbf{C}}_{-y}\mathbf{Q}_{-m} = O_p(N^{-1/2})$. Since $\mathbf{Q}_m = (0, 1)'$, we have $\widetilde{\mathbf{C}}_{-y}\mathbf{Q}_{-m} = \overline{\theta}$, which is $O_p(N^{-1/2})$ if $\overline{\theta}$ is. But we already know that $\mathbf{b}_4 = O_p(N^{-1/2})$ under this condition. Hence, in this DGP the second requirement of (iv), $\overline{\mathbf{C}}_m\mathbf{Q}_{-m} = O_p(N^{-1/2})$, is redundant. However, we note that $\overline{\mathbf{C}}_m = \mathbf{H}'\overline{\mathbf{C}}$, which under $\overline{\theta} = O_p(N^{-1})$ is asymptotically proportional to $\overline{\mathbf{C}}_y$, the first row of $\overline{\mathbf{C}}$, and therefore $\overline{\mathbf{C}}_m\mathbf{Q}_{-m}$ is asymptotically proportional to $\overline{\pi}$. One way to satisfy the second requirement of (iv) is therefore to assume that $\overline{\pi} = O_p(N^{-1/2})$.

As we just pointed out, in the DGP considered in the above example Assumption 3.1(iv) is stronger than necessary. In supplementary online appendix we provide a set of less restrictive high-level assumptions, which ensure that the rate of convergence in S4 is $\sqrt{NT}$. In terms of Example 2.2, these high-level assumptions are tantamount to requiring $\overline{\pi} = O_p(N^{-1/2})$, which will be the case if $\pi = 0$, and we have already seen that this is enough to ensure that $\mathbf{b}_4 = O_p(N^{-1/2})$. Hence, quite surprisingly, while the model to be estimated is the one in (2.1) for $y_{i,t}$, under S4 it is the loadings of the regressors in (2.2) that need to be restricted. This is true in the simple example considered here. With more than one regressor it is necessary to also restrict $\mathbf{J}_{\overline{\mathbf{E}}}$, and this is when the high-level assumptions come in.

Remark 3.2 (*Reduced Rate of Convergence in the QML Case*)**.** Consistent with our Proposition 3.2, Moon and Weidner (2015) show that the rate of convergence of the QML estimator can be reduced from $\sqrt{NT}$ to $\sqrt{\min\{N, T\}}$ whenever $r > r_y$. This reduction can be avoided, but then at the expense of additional assumptions on the errors and regressors, such as that $\boldsymbol{\varepsilon}_i$ is normally distributed and that $\mathbf{x}_i$ has certain high-order moments. Both estimation approaches therefore require additional assumptions to ensure $\sqrt{NT}$-consistency under S4. The restrictions are not the same, though, and are in fact not comparable, at least not in general.

3.3. Asymptotic distribution

As we pointed out in the previous section, the limiting behavior of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ generally depends on $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$, which enters through $\boldsymbol{\Sigma}_{\overline{\mathbf{E}}}$ in $\mathbf{b}_2$ and $\boldsymbol{\Sigma}_\mathbf{x}$. However, not in S1 and S2, as in these specifications $\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}} = \mathbf{0}_{(r-r_y)\times(r-r_y)}$, and so $\mathbf{b}_3 = \mathbf{b}_4 = \mathbf{0}_{k\times 1}$ and $\boldsymbol{\Sigma}_\mathbf{x} = \boldsymbol{\Sigma}_\mathbf{v}$. The dependence on $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$ in $\mathbf{b}_2$ is gone too, as $\boldsymbol{\Sigma}_{\overline{\mathbf{E}}} = \mathbf{D}'\boldsymbol{\Sigma}_\mathbf{e}\mathbf{D}$. This means that in S1 and S2 $\mathbf{b}_1$ and $\mathbf{b}_2$ are constant vectors. Hence, in view of Theorem 3.1, it is clear that

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) - \boldsymbol{\Sigma}_\mathbf{v}^{-1}(\sqrt{\kappa}\mathbf{b}_1 + \kappa^{-1/2}\mathbf{b}_2) = \boldsymbol{\Sigma}_\mathbf{v}^{-1}\mathbf{b}_0 + o_p(1), \tag{3.15}$$

whose asymptotic distribution is given in the next corollary to Theorem 3.1.

Corollary 3.3. *Suppose that the conditions of Theorem 3.1 hold. Then, under S1 and S2,*

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) - \boldsymbol{\Sigma}_\mathbf{v}^{-1}(\sqrt{\kappa}\mathbf{b}_1 + \kappa^{-1/2}\mathbf{b}_2) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, \boldsymbol{\Sigma}_\mathbf{v}^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_\mathbf{v}^{-1}).$$

Table 1
Consistency and asymptotic distribution.

Specification	Assumptions	Rate	Asymptotic distribution
S1	2.1–2.5	$\sqrt{NT}$	Normal
S2	2.1–2.5	$\sqrt{NT}$	Normal
S3	2.1–2.5	$\sqrt{NT}$	Normal
S4	2.1–2.5	$\sqrt{N}$	Nonstandard
S4	2.1–2.5 and 3.1	$\sqrt{NT}$	Nonstandard

Notes: "S1"–"S4" refer to the specifications with $k+1 = r_m = r$, $k+1 > r_m = r$, $k+1 = r_m < r$ and $k+1 > r_m < r$, respectively. The assumptions are given in Section 2.

Corollary 3.3 shows that under S1 and S2 the CCEP estimator is asymptotically normal. However, while normal, the asymptotic distribution of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ is not centered at zero, as to be expected given the presence of incidental parameters. In Section 3.4, we discuss the implications of this miscentering for inference.

Remark 3.3 (*Dynamics in the Heterogeneous Slope Case*)**.** Chudik and Pesaran (2015) allow for a lagged dependent variable in a DGP with heterogeneous slopes. According to their results, the heterogeneity of the slopes makes $y_{i,t}$ dependent not only on $\mathbf{F}_t$, but also on all its lags. This greatly complicates both the estimation and the underlying theory, calling for a finite-order approximation of an infinite-dimensional model. It also creates a dependence on the order of approximation, which is likely to have a substantial effect in small samples (see Chudik and Pesaran, 2015, and Chudik et al., 2017). Corollary 3.3 does not require any approximations of this kind. The reason for this much simplified theory is the equal slope assumption.

In S3, $\mathbf{J}_{\overline{\mathbf{E}}} = \mathbf{0}_{(k+1)\times(k+1)}$, which means that while $\mathbf{b}_4 = \mathbf{0}_{k\times 1}$, and $\mathbf{b}_2$ is given by Proposition 3.1. Hence, just as in S1 and S2, $\mathbf{b}_1$ and $\mathbf{b}_2$ are constant vectors. Also,

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{v}} + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\Lambda}_{-y,i}\boldsymbol{\Sigma}_{\mathbf{F}_{-m,-y}}\boldsymbol{\Lambda}'_{-y,i}, \tag{3.16}$$

which in view of Theorem 3.1 in turn implies

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) - \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\sqrt{\kappa}\mathbf{b}_1 + \kappa^{-1/2}\mathbf{b}_2) = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{b}_0 + \mathbf{b}_3) + o_p(1). \tag{3.17}$$

The asymptotic distribution of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) - \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\sqrt{\kappa}\mathbf{b}_1 + \kappa^{-1/2}\mathbf{b}_2)$ is given in Corollary 3.4.

Corollary 3.4. *Suppose that the conditions of Theorem 3.1 hold. Then, under S3,*

$$\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) - \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\sqrt{\kappa}\mathbf{b}_1 + \kappa^{-1/2}\mathbf{b}_2) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_3)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}).$$

The fact that $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ is asymptotically normal not only in S1 and S2 but also in S3 is important, because the previous literature (based on equal slopes) has not considered the case when $r > r_m$. Another interesting observation is that while the unaccounted for non-mean factors in S3 do not interfere with asymptotic normality, since $\boldsymbol{\Sigma}_3$ is positive semidefinite, they do contribute to the asymptotic variance of the CCEP estimator.

In S4, the asymptotic behavior of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ is governed by $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{b}_4$, as shown in (3.8). This term depends on $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$, which enters through both the numerator and the denominator. The fact that $\boldsymbol{\Sigma}_{\overline{\mathbf{E}}}$ is quadratic in $\sqrt{N}\,\widetilde{\mathbf{C}}_{-y}$ means that the asymptotic distribution of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ in S4 is generally not going to be normal. This is noteworthy, because it is the first time in the literature that the asymptotic distribution of the CCEP estimator has been shown to be nonstandard. It follows that, while consistent in all four specifications, asymptotic normality in general is only possible in S1–S3. Table 1 summarizes the results reported thus far on consistency and asymptotic distribution.

Example 3.1 (*Non-normality*)**.** Consider the following DGP:

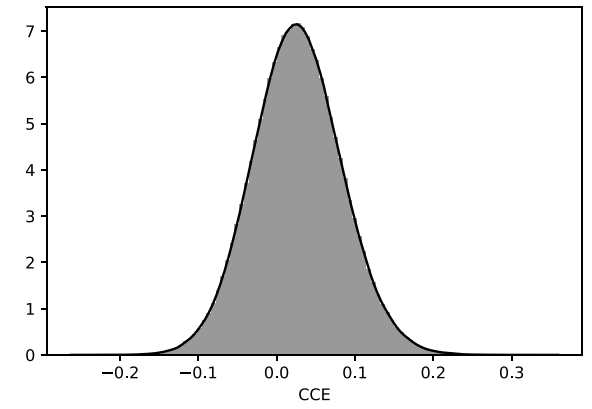
$$y_{i,t} = \beta x_{i,t} + \gamma_i F_{1,t} + \varepsilon_{i,t}, \tag{3.18}$$

$$x_{i,t} = \boldsymbol{\pi}_i'\mathbf{F}_t + v_{i,t}, \tag{3.19}$$

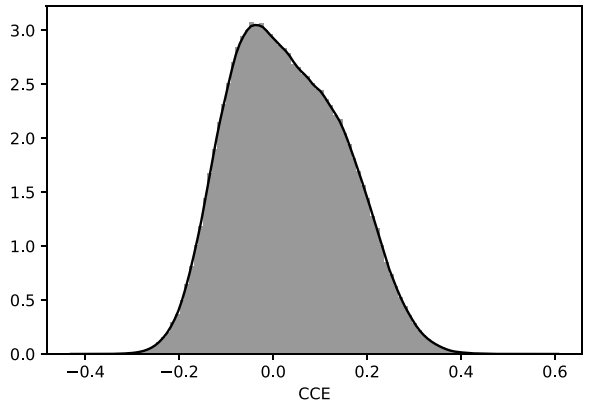
where $\boldsymbol{\pi}_i = (\gamma_i, \pi_i)$, $\mathbf{F}_t = (F_{1,t}, F_{2,t})'$ and $\beta = 0$. Note that the y-factor $F_{1,t}$ enters with the same loading in both equations. Suppose in addition that $(\mathbf{F}_t', \varepsilon_{i,t}, v_{i,t})' \sim N(\mathbf{0}_{4\times 1}, \mathbf{I}_4)$ and $(\gamma_i, \pi_i)' \sim N((1,0)', \boldsymbol{\Omega})$, where

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & \rho/\sqrt{1+\rho^2} \\ \rho/\sqrt{1+\rho^2} & 1 \end{pmatrix}. \tag{3.20}$$

Two values of ρ are considered; $\rho = 0$ and $\rho = 0.8$. If $\rho = 0$, γ_i and π_i are uncorrelated, whereas if $\rho = 0.8$, the correlation between γ_i and π_i is ≈ 0.62. Note that while $F_{1,t}$ can be well approximated by cross-sectional averages of $y_{i,t}$



(a) $\rho = 0$



(b) $\rho = 0.8$

Fig. 1. Simulated small-sample distribution of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ in S4.

and/or $x_{i,t}$, $F_{2,t}$ cannot. Under these assumptions,

$$\mathbf{b}_4 = -\frac{(N^{-1}\sum_{i=1}^{N}\gamma_i\pi_i)\sqrt{N}\overline{\pi}}{(\sqrt{N}\overline{\pi})^2 + 2}. \tag{3.21}$$

If $\rho = 0$, then $N^{-1}\sum_{i=1}^{N}\gamma_i\pi_i$ is $O_p(N^{-1/2})$ and therefore so is $\mathbf{b}_4$. Hence, in this case $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}) = O_p(1)$. Moreover, under normality $N^{-1/2}\sum_{i=1}^{N}\gamma_i\pi_i$ is asymptotically independent of $\sqrt{N}\overline{\pi}$, which means that the asymptotic distribution of $\sqrt{NT}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ is going to be mixed normal. This is true if $\rho = 0$. If $\rho \neq 0$, then $N^{-1}\sum_{i=1}^{N}\gamma_i\pi_i$ and hence also $\mathbf{b}_4$ are $O_p(1)$, and the asymptotic distribution of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ is not mixed normal, but is instead a non-linear function of $\sqrt{N}\overline{\pi} \sim N(0, 1)$. This is illustrated in Fig. 1, which reports histograms representing the simulated distribution of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})$ when $N = T = 200$. We see that while normal when $\rho = 0$, when $\rho = 0.8$ the simulated distribution exhibits marked deviations from normality.

3.4. Implications for inference

The distributional results reported in Sections 3.2 and 3.3 are important not only in their own right but also for their implications for inference. In particular, as pointed out in Section 3.3, while generally nonstandard in S4, the asymptotic distribution of the CCEP estimator in S1–S3 is normal. Of course, since the mean is not zero, valid inference relies on the availability of suitable corrections for the biases in $\mathbf{b}_1$ and $\mathbf{b}_2$. Consider $\mathbf{b}_1$, the Nickell bias. One possibility here is to employ the analytical approach of Moon and Weidner (2017), which is based on long-run variance estimation. Unfortunately, this approach only works in the most restrictive S1 specification. Another possibility is to use the split-panel jackknife (SPJ) of Dhaene and Jochmans (2015). This approach has been shown to be very effective in mitigating bias (see, for

example, Chudik and Pesaran, 2015; Galvao and Kato, 2014, and Fernández-Val and Weidner, 2016), and it is expected to work well also in the current context.

The expression for $\mathbf{b}_2$ depends on inestimable quantities such as $\mathbf{J}_{\overline{\mathbf{E}}}$, which means that analytical correction as suggested by Westerlund and Urbain (2015) will generally not work, not even in S2. The SPJ is an option. However, unlike with $\mathbf{b}_1$, SPJ of $\mathbf{b}_2$ requires dividing the cross-sectional dimension of the panel in two, which is problematic because there is no natural ordering of the cross-sectional units.[6] An alternative here is the leave-one-out jackknife (LOOJ) of Hahn and Newey (2004), which is based on N sub-samples with $N-1$ cross-sectional units in each. The LOOJ generally requires stronger assumptions than the SPJ (see Fernández-Val and Weidner, 2016), but is still expected to work, at least in S1–S3. The validity of this approach in S4 is more questionable, as in this specification the bias (and variance) will in general depend on the partition being considered. Jackknife correction is therefore only expected to work in S1–S3 when $\mathbf{b}_4 = \mathbf{0}_{k\times 1}$.

Of course, unless one of the two bias terms are known to be zero, valid inference in S1–S3 requires dealing with $\mathbf{b}_1$ and $\mathbf{b}_2$ not separately, but jointly. Two possibilities here are to follow either Fernández-Val and Weidner (2016), who suggest a double SPJ correction, or Cruz-Gonzalez et al. (2017), who propose combining the SPJ with the LOOJ. In Section 4, we use Monte Carlo simulation as a means to evaluate the effectiveness of these various bias correction approaches.

4. Monte Carlo study

4.1. Setup

In this section, we investigate the small-sample accuracy of our asymptotic results by means of Monte Carlo simulations. The DGP used for this purpose can be seen as an extended version of the one used in Example 3.1, and is given by

$$y_{i,t} = \beta x_{i,t} + \gamma_i F_{1,t} + \varepsilon_{i,t}, \tag{4.1}$$

$$x_{i,t} = \boldsymbol{\pi}_i' \mathbf{F}_t + v_{i,t} \tag{4.2}$$

where $\boldsymbol{\pi}_i = (\pi_{1,i}, \pi_{2,i}, \pi_{3,i})'$, $\mathbf{F}_t = (F_{1,t}, F_{2,t}, F_{3,t})'$ and $\beta = 2$. We generate $y_{i,t}$ and $x_{i,t}$ for $t = -50, \dots, T$ time periods and discard the first 51 observations to attenuate the effect of the initialization. The idiosyncratic errors are generated as follows:

$$v_{i,t} = \delta \epsilon_{i,t} + \zeta_{i,t}, \tag{4.3}$$

$$\epsilon_{i,t} = \alpha \epsilon_{i,t-1} + \varepsilon_{i,t-1}, \tag{4.4}$$

where $\epsilon_{i,-50} = \varepsilon_{i,-50} = 0$ and $(\varepsilon_{i,t}, \zeta_{i,t})' \sim N(\mathbf{0}_{2\times 1}, \text{diag}(1, 1-\delta^2/(1-\alpha^2)))$. The variance of $\zeta_{i,t}$ is chosen so as to ensure that $v_{i,t}$ has unit variance. To also ensure that the variance of $\zeta_{i,t}$ exists, we set $\alpha = \delta = 0.65$. This means that $x_{i,t}$ is predetermined.

Consider $\mathbf{F}_t$. According to Assumption 2.2, this vector should be stationary. In order to make sure that this is indeed the case, $\mathbf{F}_t$ is generated from the following VAR(1):

$$\mathbf{F}_t = \mathbf{B}\mathbf{F}_{t-1} + \mathbf{w}_t, \tag{4.5}$$

where $\mathbf{w}_t \sim N(\mathbf{0}_{3\times 1}, \boldsymbol{\Sigma}_{\mathbf{w}})$ with $\text{vec}\,\boldsymbol{\Sigma}_{\mathbf{w}} = (\mathbf{I}_9 - \mathbf{B}\otimes\mathbf{B})\text{vec}\,\mathbf{I}_3$ such that $\boldsymbol{\Sigma}_{\mathbf{F}} = \mathbf{I}_3$. Note that when $\mathbf{B} = \mathbf{0}_{3\times 3}$, the factors are uncorrelated over time, and so there is no Nickell bias. The DGP considered here is more interesting and sets

$$\mathbf{B} = \begin{pmatrix} 0.4 & 0.1 & 0.1 \\ 0.1 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.4 \end{pmatrix}, \tag{4.6}$$

such that the eigenvalues of $\mathbf{B}$ are bounded away from unity.

Four DGPs, henceforth denoted DGP1–DGP4, are considered, one for each of our four specifications. The difference between the four DGPs lies in how we generate the loadings. The assumption we make is that $(\gamma_i, \boldsymbol{\pi}_i')' \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)'$ and

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & 0.2 & \Omega_{13} & \Omega_{14} \\ 0.2 & 1 & 0 & 0 \\ \Omega_{13} & 0 & \Omega_{33} & 0 \\ \Omega_{14} & 0 & 0 & \Omega_{44} \end{pmatrix}. \tag{4.7}$$

We allow some of the off-diagonal elements of $\boldsymbol{\Omega}$ to be nonzero, as we believe it to be unrealistic to assume that the factor loadings in $y_{i,t}$ and $x_{i,t}$ are uncorrelated. The parameterizations of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ that we consider are summarized in Table 2.

The mean of the loadings determine the rank of $\mathbf{C}$, and by setting both the mean and variance to zero we can also exclude factors. Hence, by setting $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, we can control r_m and r. The number of observables is always given by

[6] As a solution, Fernández-Val and Weidner (2016) recommend using multiple partitions.

Table 2
Monte Carlo DGPs.

DGP	Specification	μ_1	μ_2	μ_3	μ_4	Ω_{33}	Ω_{13}	Ω_{44}	Ω_{14}	r_m	r
DGP1	S1	1	0	1	0	1	0.5	0	0	2	2
DGP2	S2	1	1	0	0	0	0	0	0	1	1
DGP3	S3	1	1	1	1	1	0.5	1	0.5	2	3
DGP4	S4	1	1	0	0	1	0.5	0	0	1	2

Notes: $(\mu_1, \ldots, \mu_4)'$ refers to the mean of $(\gamma_i, \boldsymbol{\pi}_i')'$, while Ω_{33}, Ω_{13}, Ω_{44} and Ω_{14} refer to the elements of the covariance matrix of this vector.

$k + 1 = 2$ and since the mean of γ_i is one, $r_y = 1$. In other words, by determining r_m and r, we also determine the specification to be considered. In DGP1, we set $\mu_2 = \mu_4 = \Omega_{44} = \Omega_{14} = 0$. The fact that $\mu_4 = \Omega_{44} = 0$ means that $\pi_{3,i} = 0$ for all i. This eliminates $F_{3,t}$ from the system, and therefore $r = 2$. Moreover, since

$$\mathbf{C} = \begin{pmatrix} \mu_2\beta + \mu_1 & \mu_2 \\ \mu_3\beta & \mu_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \tag{4.8}$$

has full rank, we have that $r_m = \operatorname{rk}\mathbf{C} = 2$. This is intuitive, because $\mu_2 = \mu_4 = 0$, which means that only $F_{1,t}$ and $F_{2,t}$ are permitted to affect the mean of the data. Hence, $r_m = r = k + 1 = 2$, which means that we are in S1. In DGP2, all mean and variance parameters but μ_1 and μ_2 are zero, which means that now there is just one factor ($r = 1$), $F_{1,t}$, that enters the mean of both $y_{i,t}$ and $x_{i,t}$. Hence, since

$$\mathbf{C} = (\mu_2\beta + \mu_1,\ \mu_2) = (3,\ 1), \tag{4.9}$$

we have $r_m = r = 1 < k + 1 = 2$. This DGP is therefore an example of S2. In DGP3, $\boldsymbol{\mu} = (1, \ldots, 1)'$ and all the variance parameters are positive. All three factors are included ($r = 3$) but only two affect the mean of the data, as

$$\mathbf{C} = \begin{pmatrix} \mu_2\beta + \mu_1 & \mu_2 \\ \mu_3\beta & \mu_3 \\ \mu_4\beta & \mu_4 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 2 & 1 \\ 2 & 1 \end{pmatrix} \tag{4.10}$$

has rank $r_m = 2$. Hence, $r_m = k + 1 = 2 < r = 3$, and so we are in S3. DGP4 is similar to DGP2 in that $\mathbf{C}$ is the same, and so $r_m = 1$. In this case, however, $\Omega_{33} = 1$, which means that there are $r = 2$ factors present. If follows that, $r_m = 1 < k + 1 = r = 2$, which means that we are in S4.

As discussed in Section 3.3, while consistent, valid inference based on the CCEP estimator requires bias correction. Some of the available approaches include analytical correction, SPJ and LOOJ. In this section, we consider no less than seven estimation approaches, whose only difference lies in the treatment of the bias, as described in Table 3. The effectiveness of these approaches varies, from the standard uncorrected NN approach, which is not expected to perform well in any of the specifications, to the jackknife-based SL and SS approaches that are expected to work well in S1–S3. None of the approaches are expected to work in S4, but we will still consider this specification, as in practice the DGP is never known. In addition to the bias correction, valid inference requires a consistent estimator of the asymptotic covariance matrix of the CCEP estimator. One possibility here is to follow Pesaran (2006), who proposes a simple plug-in estimator. However, this approach is only expected to work in S1 and S2. In this section, we therefore follow, for example, Galvao and Kato (2014), and Gonçalves and Kaffo (2015), and employ the cross-sectional bootstrap of Kapetanios (2008) to construct (equal-tailed) percentile confidence intervals. Kapetanios (2008) supposes that the cross-sectional units are independent, but he argues that the bootstrap should be valid also under dependence, provided that the estimator is valid, and his Monte Carlo results for the CCEP and CCEMG estimators confirm this. Based on this and the results reported in Section 3, we expect the bootstrap to perform well in S1–S3, although the performance in S4 is an open issue.

4.2. Results

We report bias, standard deviation and 5% size results over 10,000 Monte Carlo replications, where the size results are based on 499 bootstrap draws. Both the bias and the standard deviation are scaled by $\sqrt{NT}$.

We begin by discussing the results for DGP1 reported in Tables 4–6. According to Table 4, the bias of the standard CCEP estimator (NN) can be quite substantial. As expected, while the bias is roughly constant as $N = T$ increases, when $N \neq T$ the bias depends critically on the relative size of the two indices. The fact that bias generally increases (in absolute value) with N and decreases with T suggests that it is $\sqrt{\kappa}\mathbf{b}_1$, the Nickell bias, that dominates the behavior. Consistent with this we see that SPJ correction of this bias only (SN) seems to work quite well, especially among the smaller values of T, and that the effectiveness of the correction increases with N. The same is true when using analytical correction (AN). Interestingly, while smaller than for NN, the (absolute) bias of the SN and AN approaches has a tendency to grow with increases in T, which we take as a reflection of $\kappa^{-1/2}\mathbf{b}_2$, the non-Nickell bias. The correction for $\sqrt{\kappa}\mathbf{b}_1$ therefore unmasks the effect of $\kappa^{-1/2}\mathbf{b}_2$, which is otherwise dwarfed by the effect of the former bias. Among the approaches that correct for both biases, those that use SPJ correction of $\sqrt{\kappa}\mathbf{b}_1$ (SA, SL and SS) work best in terms of bias.

Table 3
CCEP estimators.

Abbreviation	Bias correction	
	$\sqrt{\kappa}\mathbf{b}_1$	$\kappa^{-1/2}\mathbf{b}_2$
NN	None	None
SN	SPJ	None
SA	SPJ	Analytical
SL	SPJ	LOOJ
AN	Analytical	None
AA	Analytical	Analytical
SS	SPJ	SPJ

Notes: "SPJ" and "LOOJ" refer to the split-panel jackknife of Dhaene and Jochmans (2015) and the leave-one-out jackknife of Hahn and Newey (2004), respectively. For the analytical corrections of $\sqrt{\kappa}\mathbf{b}_1$ and $\kappa^{-1/2}\mathbf{b}_2$ we use the formulas suggested by Moon and Weidner (2017), and Westerlund and Urbain (2015), respectively. All estimators use the cross-sectional bootstrap of Kapetanios (2008) to conduct inference.

Table 4
Bias $\times\sqrt{NT}$ for DGP1.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	−0.62	0.19	−0.13	−0.05	−0.09	−0.41	−0.03
50	20	−0.22	0.30	−0.07	−0.05	0.17	−0.21	−0.02
100	20	0.09	0.47	−0.01	−0.04	0.38	−0.10	0.00
200	20	0.40	0.68	0.04	−0.02	0.61	−0.02	0.07
20	50	−1.23	0.09	−0.12	−0.06	−0.33	−0.54	−0.06
50	50	−0.67	0.16	−0.09	−0.06	−0.01	−0.26	−0.07
100	50	−0.34	0.26	−0.05	−0.04	0.15	−0.16	−0.05
200	50	−0.02	0.41	−0.01	−0.02	0.34	−0.09	−0.03
20	100	−1.83	0.01	−0.13	−0.08	−0.53	−0.67	−0.09
50	100	−1.09	0.09	−0.09	−0.07	−0.14	−0.31	−0.07
100	100	−0.68	0.17	−0.05	−0.04	0.03	−0.19	−0.04
200	100	−0.36	0.26	−0.04	−0.04	0.16	−0.14	−0.05
20	200	−2.69	−0.04	−0.14	−0.11	−0.81	−0.91	−0.11
50	200	−1.66	0.00	−0.12	−0.10	−0.29	−0.42	−0.10
100	200	−1.08	0.11	−0.05	−0.03	−0.06	−0.22	−0.03
200	200	−0.68	0.19	−0.02	−0.01	0.07	−0.14	−0.01

Notes: See Table 2 for an explanation of DGP1–DGP4, and Table 3 for a description of the estimators considered.

Table 5
Standard deviation $\times\sqrt{NT}$ for DGP1.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	1.33	1.45	1.54	1.58	1.26	1.35	1.67
50	20	1.21	1.23	1.22	1.27	1.19	1.18	1.40
100	20	1.28	1.28	1.16	1.19	1.26	1.13	1.41
200	20	1.48	1.47	1.15	1.19	1.47	1.13	1.55
20	50	1.49	1.60	1.64	1.65	1.31	1.37	1.68
50	50	1.14	1.14	1.16	1.17	1.09	1.12	1.21
100	50	1.08	1.09	1.10	1.10	1.07	1.07	1.15
200	50	1.09	1.10	1.07	1.07	1.09	1.06	1.14
20	100	1.71	1.84	1.86	1.87	1.38	1.42	1.88
50	100	1.19	1.17	1.18	1.19	1.11	1.13	1.20
100	100	1.07	1.06	1.07	1.07	1.04	1.06	1.09
200	100	1.04	1.04	1.05	1.05	1.03	1.04	1.07
20	200	2.12	2.26	2.27	2.27	1.58	1.61	2.28
50	200	1.30	1.22	1.23	1.23	1.15	1.17	1.24
100	200	1.11	1.07	1.08	1.08	1.06	1.07	1.09
200	200	1.03	1.02	1.03	1.03	1.02	1.03	1.04

Notes: See Table 4 for an explanation.

According to Table 5, there are no major differences in terms of standard deviation, except that some estimation approaches (NN, SN, AN and SS) tend to suffer when $N = 20$, especially when T is large. However, we also see that these differences tend to disappear with increases in N, and that the standard deviations seem to be converging, as expected given the $\sqrt{NT}$-rate of convergence in S1, and the fact that the results are scaled by $\sqrt{NT}$. Similarly, when we look at Table 6, we see that, except for NN, which tend to be outperformed by the competition, the performance in terms of

Table 6
5% size for DGP1.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	5.0	4.2	3.6	2.8	3.5	3.9	1.9
50	20	2.4	4.8	2.2	1.8	4.4	2.4	1.1
100	20	3.4	4.7	2.0	1.3	4.4	1.7	0.3
200	20	2.9	3.9	1.5	1.0	3.5	1.2	0.2
20	50	13.3	7.0	6.8	7.3	5.5	6.1	6.5
50	50	4.3	4.2	3.2	3.6	4.4	3.9	2.6
100	50	3.6	5.5	4.3	3.8	4.7	3.9	2.2
200	50	2.9	6.6	3.5	3.5	6.4	3.3	1.4
20	100	30.2	10.8	9.9	10.3	10.1	11.2	9.5
50	100	12.2	5.4	5.3	5.3	5.7	5.6	5.1
100	100	6.0	6.2	5.2	5.1	4.8	4.6	4.4
200	100	5.2	6.8	4.9	4.9	6.1	5.3	4.0
20	200	48.0	18.2	18.7	18.5	13.8	15.2	18.2
50	200	26.5	7.0	7.2	7.1	6.1	6.4	6.8
100	200	14.2	5.8	6.0	5.8	6.2	6.1	5.6
200	200	5.9	6.0	5.0	5.2	5.9	5.1	4.7

Notes: See Table 4 for an explanation.

Table 7
Bias $\times\sqrt{NT}$ for DGP2.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	−0.42	0.07	−0.03	0.07	−0.12	−0.23	0.07
50	20	−0.18	0.15	−0.01	0.05	0.05	−0.11	0.05
100	20	−0.03	0.21	−0.01	0.03	0.14	−0.08	0.03
200	20	0.13	0.30	−0.01	0.01	0.25	−0.06	0.01
20	50	−0.69	0.05	0.02	0.05	−0.22	−0.25	0.05
50	50	−0.41	0.05	−0.04	−0.01	−0.05	−0.14	−0.01
100	50	−0.21	0.13	−0.01	0.02	0.06	−0.08	0.01
200	50	−0.05	0.20	0.00	0.02	0.15	−0.04	0.02
20	100	−0.98	0.03	0.02	0.01	−0.30	−0.31	0.02
50	100	−0.61	0.02	−0.02	−0.02	−0.11	−0.15	−0.02
100	100	−0.38	0.08	0.00	0.00	0.00	−0.09	0.00
200	100	−0.22	0.12	−0.01	−0.01	0.06	−0.07	−0.01
20	200	−1.35	0.05	0.10	0.04	−0.41	−0.36	0.04
50	200	−0.90	−0.02	−0.03	−0.06	−0.19	−0.19	−0.05
100	200	−0.61	0.04	−0.01	−0.02	−0.07	−0.11	−0.02
200	200	−0.37	0.10	0.02	0.01	0.03	−0.06	0.01

Notes: See Table 4 for an explanation.

size accuracy is very similar across estimation approaches. As expected, size accuracy is not perfect, and there are some distortions. Note in particular how the distortions have a tendency to accumulate and to increase with N. However, except for NN, this tendency is mainly among the smaller values of T, and there is a marked improvement in size accuracy as T increases. Note in particular how SA, SL and SS tend to perform well with only minor distortions for all panels with N and T larger than 20. Things improve also for NN, but only very slowly, which is to be expected given its relatively large bias.

Let us now consider the results reported in Tables 7–9 for DGP2. Because the variances of $\gamma_i F_{1,t} + \varepsilon_{i,t}$ and $x_{i,t}$ are allowed to vary across DGPs, the results for DGP2 are not really comparable to those reported in Tables 4–6 for DGP1. What we can say, however, is that since the asymptotic results for S1 and S2 are very similar, the relative performance of the various estimation approaches should also be similar, and this is exactly what we see when we compare the results for DGP1 and DGP2. Note in particular how the bias of SA, SL and SS tend to be relatively small (in absolute value). One difference when compared to the results reported for DGP1 is that the standard deviations of those approaches that either disregard the issue of bias completely (NN) or correct only for the Nickell bias (SN and AN) tend to be slightly smaller than the standard deviations of the other approaches, except when $T = 20$. Bias correction can therefore lead to increased variance, which is in agreement with the results reported by Fernández-Val and Weidner (2016), and Westerlund (2018).

The first thing to note about the results reported in Tables 10–12 for DGP3 is that the relative ranking of the estimation approaches in terms of bias is the same as in DGP1 and DGP2 with SA, SL and SS leading to the best performance. This accords with our a priori expectations, because bias correction should work not only in S1 and S2 but also in S3. As in DGP1 and DGP2, the difference in terms of standard deviation is small, although we see that the double bias corrected approaches (SA, SL, AA and SS) tend to perform slightly worse than the approaches based on at most one correction (NN, SN and AN). Because the relative performance in terms of bias and standard deviation is so similar to before, it is not

Table 8
Standard deviation $\times\sqrt{NT}$ for DGP2.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	1.28	1.37	1.65	1.77	1.23	1.52	1.68
50	20	1.18	1.20	1.26	1.31	1.16	1.22	1.39
100	20	1.23	1.23	1.16	1.19	1.22	1.14	1.39
200	20	1.36	1.37	1.12	1.15	1.36	1.11	1.46
20	50	1.35	1.45	1.77	1.93	1.20	1.59	1.72
50	50	1.12	1.13	1.23	1.27	1.08	1.20	1.26
100	50	1.07	1.08	1.12	1.13	1.06	1.10	1.16
200	50	1.08	1.08	1.07	1.08	1.07	1.06	1.13
20	100	1.56	1.64	2.08	2.31	1.28	1.83	1.97
50	100	1.15	1.14	1.25	1.31	1.07	1.22	1.25
100	100	1.07	1.07	1.12	1.14	1.05	1.11	1.14
200	100	1.04	1.04	1.06	1.07	1.03	1.05	1.08
20	200	1.90	1.95	2.55	2.90	1.41	2.20	2.37
50	200	1.23	1.18	1.34	1.42	1.09	1.30	1.31
100	200	1.08	1.06	1.11	1.14	1.04	1.11	1.12
200	200	1.03	1.03	1.05	1.06	1.02	1.04	1.06

Notes: See Table 4 for an explanation.

Table 9
5% size for DGP2.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	3.9	1.9	1.0	0.3	2.8	1.4	0.4
50	20	3.6	4.9	2.3	1.8	4.6	2.1	0.6
100	20	3.8	5.2	2.4	1.3	4.7	2.6	0.5
200	20	4.7	7.0	2.1	0.6	6.9	1.6	0.0
20	50	10.1	1.6	1.7	0.9	3.5	3.2	0.8
50	50	7.3	3.7	3.0	1.9	3.9	4.2	1.6
100	50	4.5	7.4	3.9	3.4	5.9	3.9	2.6
200	50	4.5	8.8	4.0	4.6	8.3	3.8	2.0
20	100	22.0	2.1	3.0	1.2	4.7	5.9	0.9
50	100	14.2	3.8	3.5	1.9	3.6	5.0	1.8
100	100	7.5	4.4	3.4	3.4	4.9	4.7	2.8
200	100	4.9	7.9	5.4	4.8	6.9	5.0	4.1
20	200	38.8	2.2	3.9	1.7	7.3	10.4	1.6
50	200	30.3	3.6	5.9	2.3	4.7	10.2	2.1
100	200	15.0	4.3	4.7	2.9	4.3	6.1	2.5
200	200	7.2	4.9	4.2	3.4	3.7	4.7	3.0

Notes: See Table 4 for an explanation.

surprising to see that conclusions regarding the size results are largely the same. The fact that the size distortions tend to disappear with increases in T is consistent with the asymptotic normality of Corollary 3.4 suggesting that asymptotically size accuracy should be perfect.

In DGP4, due to the presence of $\sqrt{T}\mathbf{b}_4$, the rate of convergence is reduced from $\sqrt{NT}$ to $\sqrt{N}$. This is reflected in the results reported in Tables 13–15. Note in particular how the standard deviations of all the estimation approaches have a tendency to increase (in absolute terms) with T, which is to be expected given the $\sqrt{N}$-rate of convergence and the scaling by $\sqrt{NT}$. In order isolate the effect of $\sqrt{T}\mathbf{b}_4$, we look at the panel constellations where $N = T$. We see that a doubling of the sample size generally leads to a less-than-double increase in standard deviation, which is in agreement with the $\sqrt{T}$-rate of divergence in this case. In terms of bias, SA is not performing as well as before, which is not unexpected as the analytical correction is no longer addressing the dominating bias term. SL and SS are also not addressing the dominating bias term, but unlike analytical correction, these approaches do not make any assumptions regarding the exact form of the bias, and it seems as that they are quite effective in mopping up the bias also in S4.

The first thing to note about the size results reported in Table 15 is that all estimators are size distorted, and that there is generally no improvement as N and T increase. The bootstrap is therefore unable to cope with DGP4, which is partly expected given that the asymptotic distribution of the CCEP estimator is non-normal in S4, and the requirement of Kapetanios (2008) that the estimator should be valid. However, we also see that, unlike in DGP1–DGP3 where the estimators are generally oversized, in DGP4 the distortions go in the other direction, leading to conservative tests. This is true not only in the particular DGP considered here, but in all simulations under S4 that we performed. Focusing again on the best-performing SL and SS approaches, we see that the sizes reported in Table 15 range from 0.3% to 1.8%. If this

Table 10
Bias $\times\sqrt{NT}$ for DGP3.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	−0.58	−0.18	−0.09	−0.03	−0.32	−0.23	−0.06
50	20	−0.52	−0.25	−0.10	−0.03	−0.32	−0.17	−0.07
100	20	−0.51	−0.32	−0.08	−0.01	−0.37	−0.13	−0.07
200	20	−0.59	−0.45	−0.11	−0.01	−0.48	−0.14	−0.11
20	50	−0.77	−0.12	−0.04	−0.01	−0.33	−0.26	−0.01
50	50	−0.58	−0.17	−0.04	−0.02	−0.25	−0.13	−0.02
100	50	−0.52	−0.21	−0.03	0.00	−0.27	−0.09	−0.01
200	50	−0.51	−0.29	−0.03	0.00	−0.33	−0.07	0.01
20	100	−1.02	−0.12	−0.07	−0.05	−0.39	−0.34	−0.05
50	100	−0.72	−0.13	−0.04	−0.03	−0.25	−0.16	−0.02
100	100	−0.59	−0.16	−0.03	−0.01	−0.23	−0.10	−0.01
200	100	−0.50	−0.19	−0.01	0.01	−0.24	−0.06	0.01
20	200	−1.37	−0.13	−0.10	−0.08	−0.48	−0.45	−0.08
50	200	−0.96	−0.13	−0.06	−0.05	−0.29	−0.22	−0.05
100	200	−0.72	−0.12	−0.03	−0.02	−0.22	−0.12	−0.02
200	200	−0.58	−0.15	−0.02	−0.01	−0.22	−0.08	−0.01

Notes: See Table 4 for an explanation.

Table 11
Standard deviation $\times\sqrt{NT}$ for DGP3.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	1.27	1.37	1.51	1.55	1.22	1.39	1.57
50	20	1.14	1.16	1.21	1.23	1.12	1.17	1.30
100	20	1.11	1.12	1.12	1.14	1.11	1.11	1.24
200	20	1.14	1.15	1.10	1.11	1.14	1.09	1.24
20	50	1.37	1.47	1.64	1.66	1.23	1.45	1.61
50	50	1.12	1.13	1.18	1.20	1.09	1.15	1.21
100	50	1.06	1.07	1.09	1.10	1.04	1.07	1.12
200	50	1.05	1.06	1.06	1.07	1.05	1.06	1.10
20	100	1.54	1.64	1.87	1.88	1.29	1.60	1.78
50	100	1.14	1.12	1.18	1.20	1.08	1.15	1.18
100	100	1.07	1.07	1.09	1.10	1.05	1.08	1.11
200	100	1.04	1.04	1.05	1.05	1.03	1.04	1.07
20	200	1.86	1.95	2.25	2.25	1.42	1.87	2.10
50	200	1.21	1.16	1.24	1.25	1.10	1.20	1.22
100	200	1.07	1.06	1.09	1.09	1.04	1.07	1.09
200	200	1.03	1.03	1.04	1.04	1.02	1.03	1.04

Notes: See Table 4 for an explanation.

degree of size distortion is acceptable, then it would appear as that the CCEP estimator can be used not only for estimation but also for inference.

All-in-all, we find that the asymptotic theory provided in this paper provides an accurate guide to the small-sample behavior of the CCEP estimator. The main finding is that if we want to entertain the possibility that some of the factors might not be estimable ($r_m < r$), then the usefulness of the estimator depends on whether $k+1 = r_m$ or $k+1 > r_m$. On the one hand, if $k+1 = r_m$, so that the number of estimated factors (cross-section averages) is equal to the number of mean factors, then it is possible to correct for bias and to conduct accurate inference, provided that T is not too small. If, on the other hand, $k+1 > r_m$, so that the number of mean factors is overspecified, although double correction based on SPJ and LOOJ seems to work quite well in eliminating the bias, (bootstrap) inference do tend to lead to conservative tests.

5. Concluding remarks

This paper studies the properties of the CCEP estimator when the standard rank condition of Pesaran (2006) for $\sqrt{NT}$-consistency and asymptotic normality, $k+1 \geq r$, is violated. As a starting point we take a DGP that is very general, and that includes many existing DGPs as special cases. In particular, the rank condition that we consider is quite general and allows for r factors, out of which r_m are estimable when using the cross-sectional averages of the observables. The assumption we make is that $r_m \leq \min\{r, k+1\}$, which allows r to be larger than $k+1$, provided that $k+1 \geq r_m$. This seems reasonable. Indeed, while economic theory is often suggestive of only a small number of important factors (see, for example, Eberhardt et al., 2013), we cannot rule out the possibility that there are also other, less important factors, whose number might exceed the number of observables.

Table 12
5% size for DGP3.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	10.6	5.2	4.9	4.3	6.2	5.5	2.9
50	20	9.3	5.5	3.8	3.1	6.3	5.0	1.7
100	20	6.8	4.6	3.0	2.7	5.5	2.6	1.0
200	20	6.9	5.1	2.4	1.8	5.4	2.7	0.7
20	50	24.1	8.6	7.7	7.5	10.1	9.1	6.9
50	50	15.3	7.2	6.6	6.7	9.0	7.8	5.9
100	50	12.7	6.7	4.6	4.1	8.2	5.8	3.3
200	50	13.3	7.3	5.0	4.9	8.4	5.5	2.7
20	100	31.6	12.3	11.3	11.2	12.5	11.7	10.1
50	100	22.6	6.4	5.7	5.7	9.6	7.1	5.3
100	100	17.6	7.4	6.6	6.4	9.4	7.7	5.8
200	100	15.8	8.1	6.1	5.8	9.0	6.9	4.9
20	200	46.1	18.0	17.7	17.8	15.2	14.8	17.6
50	200	32.4	7.2	7.1	7.0	8.8	8.4	6.5
100	200	21.6	7.2	5.8	5.6	8.4	7.1	5.5
200	200	18.5	7.0	5.5	5.2	8.0	6.1	4.5

Notes: See Table 4 for an explanation.

Table 13
Bias $\times\sqrt{NT}$ for DGP4.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	−0.16	0.15	−0.07	0.03	0.04	−0.17	0.04
50	20	0.02	0.22	0.00	0.00	0.17	−0.05	0.01
100	20	0.19	0.33	0.07	0.01	0.30	0.04	0.02
200	20	0.35	0.46	0.14	−0.01	0.44	0.12	0.02
20	50	−0.38	0.08	−0.16	0.00	−0.06	−0.29	0.00
50	50	−0.15	0.15	−0.07	−0.01	0.09	−0.13	0.00
100	50	−0.01	0.21	0.01	0.01	0.17	−0.03	0.01
200	50	0.14	0.30	0.06	0.00	0.27	0.03	0.00
20	100	−0.61	0.01	−0.28	−0.06	−0.16	−0.45	−0.04
50	100	−0.33	0.08	−0.13	−0.02	0.01	−0.21	−0.02
100	100	−0.16	0.14	−0.06	−0.01	0.09	−0.11	0.00
200	100	−0.02	0.20	−0.02	−0.03	0.16	−0.06	−0.03
20	200	−0.90	−0.02	−0.39	−0.09	−0.26	−0.63	−0.07
50	200	−0.53	0.04	−0.22	−0.03	−0.05	−0.31	−0.02
100	200	−0.32	0.10	−0.11	−0.01	0.04	−0.17	−0.01
200	200	−0.16	0.14	−0.04	0.01	0.10	−0.09	0.00

Notes: See Table 4 for an explanation.

Table 14
Standard deviation $\times\sqrt{NT}$ for DGP4.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	1.08	1.15	1.45	1.47	1.05	1.39	1.40
50	20	1.17	1.20	1.40	1.45	1.18	1.39	1.44
100	20	1.42	1.45	1.59	1.65	1.44	1.59	1.68
200	20	1.84	1.86	1.97	2.06	1.86	1.97	2.15
20	50	1.13	1.13	1.48	1.49	1.01	1.42	1.34
50	50	1.12	1.11	1.32	1.36	1.10	1.32	1.28
100	50	1.27	1.29	1.46	1.53	1.28	1.46	1.45
200	50	1.63	1.65	1.79	1.90	1.64	1.79	1.83
20	100	1.31	1.21	1.62	1.62	1.05	1.55	1.42
50	100	1.15	1.09	1.33	1.38	1.07	1.33	1.26
100	100	1.27	1.26	1.45	1.53	1.26	1.45	1.43
200	100	1.57	1.58	1.74	1.87	1.58	1.75	1.76
20	200	1.60	1.36	1.86	1.86	1.11	1.77	1.60
50	200	1.26	1.10	1.36	1.41	1.08	1.38	1.27
100	200	1.28	1.23	1.41	1.49	1.22	1.42	1.37
200	200	1.57	1.55	1.72	1.86	1.56	1.72	1.73

Notes: See Table 4 for an explanation.

Table 15
Rejection frequencies for DGP4.

T	N	NN	SN	SA	SL	AN	AA	SS
20	20	3.0	2.3	1.2	0.8	1.8	1.1	0.6
50	20	2.3	3.8	1.3	0.4	3.8	0.8	0.2
100	20	3.5	4.2	1.9	0.6	4.3	1.8	0.5
200	20	3.2	4.2	1.8	0.3	3.7	1.5	0.3
20	50	3.4	2.5	1.5	0.9	2.7	1.5	0.8
50	50	2.9	4.3	1.6	0.9	3.9	2.0	0.8
100	50	3.2	3.3	2.0	0.9	3.4	2.2	0.9
200	50	3.0	3.8	2.6	1.0	3.9	2.4	0.9
20	100	6.6	1.8	0.9	0.8	3.9	1.9	0.7
50	100	3.7	3.1	1.6	1.2	3.2	1.4	1.0
100	100	3.1	3.7	2.4	1.8	3.6	2.6	1.5
200	100	2.8	4.5	3.4	1.5	4.1	2.8	1.6
20	200	8.4	1.2	0.7	0.8	4.3	2.2	1.1
50	200	4.2	2.5	1.4	1.2	4.7	2.3	1.3
100	200	2.8	1.9	2.0	1.8	4.9	2.6	1.4
200	200	1.7	1.2	2.2	1.5	5.2	3.4	1.2

Notes: See Table 4 for an explanation.

We find that the consistency of the CCEP estimator seem quite robust to violations of the standard rank condition. In particular, while consistency holds generally under the new condition, for asymptotic normality to be possible we have to assume that either $k+1 = r_m$, so that the number of estimated factors is equal to the true number of mean factors, or $r_m = r$, so that all the factors are estimable. The small-sample implications of these asymptotic results are investigated by means of Monte Carlo simulation. The main conclusion is that the stated conditions on k, r_m and r are important in determining the small-sample performance of the various (bias corrected) versions of the CCEP estimator that we consider, and hence that they cannot be ignored. The case when $r_m < \min\{r, k+1\}$ turns out to be particularly problematic, which is just as expected given the non-normality of the estimator. However, the size distortions do go in the "right" direction, leading to conservative tests (at least for the setups we consider). Hence, if such tests are deemed acceptable, the CCEP estimator can be used for estimation and inference in general under the new rank condition.

On the other hand, if the conservative tests are unacceptable, then it is important to be able to rule out the case when $r_m < \min\{r, k+1\}$. One possibility here is to subject the CCEP residuals to a test for cross-sectional correlation, e.g. the CD test of Pesaran (2004). However, as shown recently by Juodis and Reese (2018), while intuitively appealing, this approach suffers from the incidental parameters problem that cannot be easily fixed by means of simple bias-correction. Given the complexity of the aforementioned problem, we leave it to be investigated in the future.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.06.002.

References

Ahn, S.C., Lee, Y.H., Schmidt, P., 2013. Panel data models with multiple time-varying individual effects. J. Econometrics 174 (1), 1–14.
Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77 (4), 1229–1279.
Cesa-Bianchi, A., Pesaran, M.H., Rebucci, A., 2019. Uncertainty and economic activity: A multicountry perspective. Rev. Financ. Stud. (forthcoming).
Chudik, A., Mohaddes, K., Pesaran, M.H., 2017. Is there a debt-threshold effect on output growth?. Rev. Econ. Stat. 99 (1), 135–150.
Chudik, A., Pesaran, M.H., 2015. 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., Tosetti, E., 2011. Weak and strong cross-section dependence and estimation of large panels. Econom. J. 14 (1), 45–90.
Cruz-Gonzalez, M., Fernandez-Val, I., Weidner, M., 2017. Bias corrections for probit and logit models with two-way fixed effects. Stata J. 17 (3), 517–545.
Dhaene, G., Jochmans, K., 2015. Split-panel jackknife estimation of fixed-effect models. Rev. Econom. Stud. 82 (3), 991–1030.
Eberhardt, M., Helmers, C., Strauss, H., 2013. Do spillovers matter when estimating private returns to R&D?. Rev. Econ. Stat. 95, 436–448.
Eberhardt, M., Teal, F., 2020. The magnitude of the task ahead: macro implications of heterogeneous technology. Review of Income and Wealth 66 (2), 334–360.
Everaert, G., De Groote, T., 2016. Common correlated effects estimation of dynamic panels with cross-sectional dependence. Econometric Rev. 35, 428–463.
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.
Gonçalves, S., Kaffo, M., 2015. Bootstrap inference for linear dynamic panel data models with individual fixed effects. J. Econometrics 186 (2), 407–426.
Greenaway-McGrevy, R., Han, C., Sul, D., 2012. Asymptotic distribution of factor augmented estimators for panel regression. J. Econometrics 169 (1), 48–53.
Hahn, J., Newey, W., 2004. Jackknife and analytical Bias reduction for nonlinear panel models. Econometrica 72, 1295–1319.
Holtz-Eakin, D., Newey, W.K., Rosen, H.S., 1988. Estimating vector autoregressions with panel data. Econometrica 56, 1371–1395.

Juodis, A., Reese, S., 2018. The incidental parameters problem in testing for remaining cross-section correlation. arXiv e-prints, arXiv:1810.03715.
Juodis, A., Sarafidis, V., 2018. Fixed t dynamic panel data estimators with multi-factor errors. Econometric Rev. 37 (8), 893–929.
Kapetanios, G., 2008. A bootstrap procedure for panel data sets with many cross-sectional units. Econom. J. 11 (2), 377–395.
Kapetanios, G., Pesaran, M.H., Yamagata, T., 2011. Panels with non-stationary multifactor error structures. J. Econometrics 160 (2), 326–348.
Karabiyik, H., Reese, S., Westerlund, J., 2017. On the role of the rank condition in CCE estimation of Factor-augmented panel regressions. J. Econometrics 197 (1), 60–64.
Moon, H.R., Weidner, M., 2015. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica 83 (4), 1543–1579.
Moon, H.R., Weidner, M., 2017. Dynamic linear panel regression models with interactive fixed effects. Econometric Theory 33 (1), 158–195.
Neyman, J., Scott, E.L., 1948. Consistent estimation from partially consistent observations. Econometrica 16, 1–32.
Nickell, S., 1981. Biases in dynamic models with fixed effects. Econometrica 49, 1417–1426.
Pesaran, M.H., 2004. General Diagnostic Tests fro Cross Section Dependence in Panels. CESifo Working Paper No. 1229.
Pesaran, M.H., 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74 (4), 967–1012.
Pesaran, M.H., 2007. A simple panel unit root test in the presence of cross section dependence. J. Appl. Econometrics 22, 265–312.
Robertson, D., Sarafidis, V., 2015. IV estimation of panels with factor residuals. J. Econometrics 185 (2), 526–541.
Westerlund, J., 2015. The effect of recursive detrending on panel unit root tests. J. Econometrics 185, 453–467.
Westerlund, J., 2018. CCE in panels with general unknown factors. Econom. J. 21 (3), 264–276.
Westerlund, J., Urbain, J.-P., 2013. On the estimation and inference in factor-augmented panel regressions with correlated loadings. Econom. Lett. 119 (3), 247–250.
Westerlund, J., Urbain, J.-P., 2015. Cross-sectional averages versus principal components. J. Econometrics 185 (2), 372–377.