Alternative estimation approaches for the factor augmented panel data model with small T

Jörg Breitung*
University of Cologne

Philipp Hansen
University of Cologne

July 9, 2020

Abstract

In this paper we compare alternative estimation approaches for factor augmented panel data models. Our focus lies on panel data sets where the number of panel groups (N) is large relative to the number of time periods (T). The Principal Component (PC) and Common Correlated Effects (CCE) estimators were originally developed for panel data with large N and T, whereas the GMM approaches of Ahn et al. (2013) and Robertson and Sarafidis (2015) assumes that T is small (that is T is fixed in the asymptotic analysis). Our comparison of existing methods addresses three different issues. First, we analyze the possibility of an inappropriate normalization of the factor space (the so-called normalization failure). In particular we propose a variant of the CCE estimator that avoids the normalization failure by adapting a weighting scheme inspired by the analysis of Mundlak (1978). Second, we demonstrate how the design of the Monte Carlo simulations favors some estimators, which explains the conflicting findings from existing Monte Carlo experiments. Third, we analyze the effects of estimating versus fixing the number of factors in advance.

JEL classification: C23, C38

Keywords: Panel data, interactive fixed effects, CCE estimator, GMM

*Corresponding author: Jörg Breitung: University of Cologne, Institute of Econometrics, Albertus Magnus Platz, 50923 Köln, Germany, email: breitung@statistik.uni-koeln.de.

1 Introduction

The seminal work of Holtz-Eakin et al. (1988) has provided two important contributions to the statistical analysis of panel data. First, it proposes a GMM framework for estimating dynamic panel data models that were further developed and popularized by Arellano and Bond (1991). This approach has become standard in the dynamic analysis of panel data. The second contribution, the introduction of time varying individual effects, was less influential and went largely unnoticed for many years. For example, the excellent monograph of Baltagi (2005) – as all other textbooks on panel data analysis of the early 2000s – does not consider time varying individual effects or any other factor structure. Bai (2009) pointed out that time varying individual effects are just a special case of a factor structure and provided a general framework for estimating a panel data model with "interactive fixed effects", which is also referred to as the *factor-augmented panel data model*.

With the work of Ahn et al. (2001, 2013), Pesaran (2006), and Bai (2009) the interest in models that account for time varying heterogeneity and cross-section correlation surged considerably and the 25th International Conference on Panel Data in Vilnius 2019 included a large number of papers dealing with factor-augmented panel data models. In empirical practice, the Common Correlated Effects (CCE) approach proposed by Pesaran (2006) has recently become very popular among empirical researchers. This is due to the fact that this estimator is easy to understand and implement, a STATA routine (xtmg) and a Gretl add-on (xtcsd) is available and it performs well in Monte Carlo studies. It is however not clear, whether the CCE approach is similarly attractive in empirical applications where the number of time periods T is small (say 5 – 15). Ahn et al. (2013) and Robertson and Sarafidis (2015) proposed a GMM approach that is shown to be consistent for finite T, whereas the CCE and the Principal Component (PC) estimator were developed for samples with large T and N. Su and Jin (2012) and Westerlund et al. (2019) showed that the CCE approach is consistent and asymptotically (mixed) normal if T is fixed and $N \to \infty$, whereas the consistency of the PC estimator requires quite restrictive assumptions (such as i.i.d. errors across time) in this case. It is however not clear how large T should be in order to ensure reliable estimation and inference.

An important assumption for the CCE estimator is that the (weighted) mean

of the factor loadings is different from zero. This assumption is difficult to verify as the factors loadings are typically unknown. Furthermore, we show that the CCE estimator is already biased if the mean of the factor loadings is $O(N^{-1/2})$. To escape such a "normalization failure", we suggest a data dependent weighting scheme that is inspired by the Mundlak (1978) approach. In our Monte Carlo simulations we show that this simple weighting scheme performs well, whenever the original CCE estimator suffers from a normalization failure.

The rest of the paper is organized as follows. Section 2 compares the existing estimation methods and Section 3 reviews and complements the asymptotic results for fixed T and $N \to \infty$. Possible problems with the normalization of the estimators are analyzed in Section 4. An extension to multiple factors is considered in Section 5 and empirical approaches for selecting the number of common factors are examined in Section 6. We argue that popular selection rules for the number of factors are generally inconsistent if T is fixed. The small sample properties of alternative estimation procedures are investigated in Section 7. Specifically, we illustrate the detrimental effect of a normalization failure and demonstrate the robustness of the Mundlak type CCE estimator. Furthermore, we investigate the effects of estimating the number of factors on the performance of the estimation procedures. Finally, we employ three general model setups from the literature in order to compare the competing methods in more challenging and realistic scenarios. Section 8 concludes.

2 Existing estimation approaches

Consider the factor augmented panel data model:[1]

$$y_{it} = \boldsymbol{\beta}' \boldsymbol{x}_{it} + e_{it} \tag{1}$$

$$\text{with} \quad e_{it} = \lambda_i f_t + u_{it} \ , \tag{2}$$

where $\boldsymbol{x}_{it}$ and $\boldsymbol{\beta}$ are $k \times 1$ vectors. For the ease of exposition we first consider a single factor with $r = 1$, that is, f_t and λ_i are scalars. The extension to multiple factors is considered in Section 5.

[1]The model may include further terms such as $\boldsymbol{\gamma}_i' \boldsymbol{d}_t$, where $\boldsymbol{d}_t$ is some observed strictly exogenous regressor, cf. Pesaran (2006). As such additional terms are easily accounted for without affecting the main results, these extensions are ignored.

We adopt a "classical" panel data framework where the coefficient vector $\boldsymbol{\beta}$ is the same for all cross-section units (homogeneous panel). Furthermore, we assume that T may be small relative to N, which is typical for many panel data applications. It should be noted that the asymptotic framework of Pesaran (2006) and Bai (2009) assumes that N and T tend to infinity, whereas Ahn et al. (2013) and Robertson and Sarafidis (2015) suppose that T is small and fixed. Furthermore, the latter approach treats f_t as parameters and thereby avoids making any assumptions on these parameters, whereas Pesaran (2006) and Bai (2009) assume that the factors are weakly correlated random variables and the loadings are treated as parameters (or also as random variables). We make the assumption that u_{it} is independent (strictly exogenous) of $\boldsymbol{x}_{it}$, f_t and λ_i. This rules out dynamic specifications.[2]

It is well known that in the two way panel data model the individual and time specific effects (which result as special cases of the factor model with constant factor and loading, respectively) can be removed by a simple data transformation, where the variables are adjusted by the individual and time specific averages. It is not difficult to see that a similar transformation exists for the model with interactive fixed effects, which is given by

$$y_{it} - \lambda_i \overline{y}_t(\boldsymbol{\lambda}) = \boldsymbol{\beta}' \left[\boldsymbol{x}_{it} - \lambda_i \overline{\boldsymbol{x}}_t(\boldsymbol{\lambda})\right] + u_{it} - \lambda_i \overline{u}_t(\boldsymbol{\lambda}), \tag{3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$ and

$$\overline{y}_t(\boldsymbol{\lambda}) = \frac{1}{N\overline{\lambda^2}} \sum_{i=1}^{N} \lambda_i y_{it}$$

with $\overline{\lambda^2} = N^{-1} \sum_{i=1}^{N} \lambda_i^2$. The weighted averages $\overline{\boldsymbol{x}}_t(\boldsymbol{\lambda})$ and $\overline{u}_t(\boldsymbol{\lambda})$ are constructed in an analogous manner. Note that $\overline{e}_t(\boldsymbol{\lambda}) = \overline{y}_t(\boldsymbol{\lambda}) - \boldsymbol{\beta}'\overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}) = f_t + \overline{u}_t(\boldsymbol{\lambda})$ serves as an estimate of f_t. Estimating the transformed regression (3) is equivalent to the least-squares estimator, treating $\boldsymbol{\beta}$ and $f_1, \ldots, f_T$ as parameters and $\boldsymbol{x}_{it}$ and λ_i as regressors. Accordingly, the resulting estimator is efficient if $u_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

[2]In panels with individual specific parameters and fixed T, including weakly dependent regressors (such as lagged dependent variables) result in a bias of order $1/T$ (the incidental parameter problem). The GMM based estimators of Section 2.3 are able to cope with this bias by introducing time dependent vectors of instruments. In this paper we abstract from such complications. The reader interested in dynamic models is referred to Juodis and Sarafidis (2018).

2.1 The PC estimator

For the PC approach suggested by Bai (2009), equation (3) is replaced by the feasible version

$$y_{it} - \widehat{\lambda}_i \overline{y}_t(\widehat{\boldsymbol{\lambda}}) = \boldsymbol{\beta}' \left[\boldsymbol{x}_{it} - \widehat{\lambda}_i \overline{\boldsymbol{x}}_t(\widehat{\boldsymbol{\lambda}}) \right] + e_{it} - \widehat{\lambda}_i \overline{e}_t(\widehat{\boldsymbol{\lambda}}), \tag{4}$$

where $e_{it} = y_{it} - \boldsymbol{\beta}'\boldsymbol{x}_{it} = \lambda_i f_t + u_{it}$ and $\widehat{\lambda}_i$ denotes the PC estimator of the factor loading λ_i, which is equivalent to the eigenvector associated with the largest eigenvalue of the sample covariance matrix $\boldsymbol{\Omega}_{ee}(\boldsymbol{\beta}) = T^{-1} \sum_{t=1}^{T} \boldsymbol{e}_t(\boldsymbol{\beta}) \boldsymbol{e}_t(\boldsymbol{\beta})'$ with $\boldsymbol{e}_t(\boldsymbol{\beta}) = (y_{i1} - \boldsymbol{\beta}'\boldsymbol{x}_{i1}, \ldots, y_{iT} - \boldsymbol{\beta}'\boldsymbol{x}_{iT})'$. As shown by Moon and Weidner (2015) the sum of squared residuals can be obtained by minimizing the objective function

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin} \left\{ \mu_{\min} \left[\sum_{i=1}^{N} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} \right) \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} \right)' \right] \right\} \tag{5}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$ and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})'$ and $\mu_{\min}\{\boldsymbol{A}\}$ denotes the smallest eigenvalue of the matrix $\boldsymbol{A}$. The minimum can be obtained by standard numerical methods, whereas Bai (2009) proposed to compute the (nonlinear) least-squares estimator of (4) sequentially by starting with the pooled OLS or within-group estimator of $\boldsymbol{\beta}$ (that is by ignoring the factor structure in the errors). The first principal component of the residual $e_{it}(\widehat{\boldsymbol{\beta}})$ yields a first estimator of the common factor and the associated loadings are used to obtain the estimated analog of the weighted averages in (4). The estimation procedure is iterated until the estimators converge to the least-squares estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$.

Moon and Weidner (2019) pointed out that the least-squares objective function may exhibit several local minima and therefore it is possible that the gradient based minimization based algorithm fails to find the global minimum. To cope with this problem, Moon and Weidner (2019) propose a nuclear norm penalty that results in a convex optimization problem. Another possibility is to initialize the minimization algorithm by a $\sqrt{NT}$-consistent initial estimator. In this case it is sufficient to assume convexity in the $1/\sqrt{NT}$ vicinity around the true value.

2.2 The CCE Estimator

In contrast to the PC estimator, the CCE approach proposed by Pesaran (2006) does not adopt an (asymptotically) efficient weighting scheme, but employs instead pre-specified weights $\boldsymbol{\lambda}_0$.[3] In practice $\boldsymbol{\lambda}_0 = (1, \ldots, 1)'$ is the default option, but any other granual weighting scheme is possible. This gives rise to a modified transformation,

$$y_{it} - \lambda_i^* \, \overline{y}_t(\boldsymbol{\lambda}_0) = \boldsymbol{\beta}' \left[\boldsymbol{x}_{it} - \lambda_i^* \, \overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}_0) \right] + u_{it} - \lambda_i^* \, \overline{u}_t(\boldsymbol{\lambda}_0), \tag{6}$$

where

$$\lambda_i^* = \lambda_i \frac{\sum_{i=1}^N \lambda_{0,i}^2}{\sum_{i=1}^N \lambda_{0,i} \lambda_i}$$

is required to drop the factor from the model. Note that if $\lambda_{0,i} = \lambda_i$ for all i, then $\lambda_i^* = \lambda_i$ and the transformation is equivalent to (3). Furthermore, if $\lambda_{0,i} = 1$ then $\lambda_i^* = \lambda_i / \overline{\lambda}$, where $\overline{\lambda} = N^{-1} \sum_{i=1}^N \lambda_i$. By reorganizing (6), we obtain the cross-section augmented regression equation,

$$y_{it} = \boldsymbol{\beta}' \boldsymbol{x}_{it} + \lambda_i^* \overline{y}_t(\boldsymbol{\lambda}_0) + \boldsymbol{\gamma}_i' \overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}_0) + v_{it} \ , \tag{7}$$

where $\boldsymbol{\gamma}_i = -\lambda_i^* \boldsymbol{\beta}$ and $v_{it} = e_{it} - \lambda_i^* \overline{e}_t(\boldsymbol{\lambda}_0)$. In practice, the nonlinear restriction $\boldsymbol{\gamma}_i = -\lambda_i^* \boldsymbol{\beta}$ is ignored and, therefore, $\boldsymbol{\gamma}_i$ is treated as an additional parameter.[4]

2.3 The HNR and ALS approach

While the CCE and PC approach replace the unobserved *factor* by (weighted) averages of $y_{1t}, \ldots, y_{Nt}$ and $\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Nt}$, the approaches suggested by Holtz-Eakin et al. (1988) (HNR) and Ahn et al. (2013) (ALS) replace the unknown factor

[3] This does not imply, however, that the CCE estimator is always inefficient whenever $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}_0$. As shown by Westerlund et al. (2019) the CCE estimator is asymptotically efficient if $r = k+1$ and u_{it} is i.i.d. across i and t.

[4] The restricted version of the CCE estimator is considered in Everaert and De Groote (2016). In our experience, imposing the nonlinear restriction does not result in an important gain in efficiency. In the model with $r > 1$ the restriction cannot be imposed anyway.

loadings by linear combinations of $y_{i1}, \ldots, y_{iT}$ and $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}$:

$$\text{HNR:} \quad \frac{1}{f_{t-1}}(y_{i,t-1} - \boldsymbol{\beta}'\boldsymbol{x}_{i,t-1}) \;=\; \lambda_i + \frac{1}{f_{t-1}}u_{i,t-1} \tag{8}$$

$$\text{ALS:} \quad \frac{1}{f_T}(y_{iT} - \boldsymbol{\beta}'\boldsymbol{x}_{iT}) \;=\; \lambda_i + \frac{1}{f_T}u_{iT}\ . \tag{9}$$

The main difference between these two approaches is that in (8) the linear combination is time dependent whereas in (9) the linear combination is the same for all time series. As we do not see any advantage in using the variant HNR (and in our simulations the HNR estimator tends to perform worse than the ALS estimator), we focus on the ALS variant in the following analysis.

Inserting (9) in the model (1) yields

$$\text{ALS:} \quad y_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \theta_t y_{iT} - \theta_t\boldsymbol{\beta}'\boldsymbol{x}_{iT} + \nu_{it} \qquad \text{for } t = 1, \ldots, T-1, \tag{10}$$

where $\theta_t = f_t/f_T$ and $\nu_{it} = u_{it} - \theta_t u_{i,T}$. Note that this approach involves $T-1$ additional parameters $\theta_1, \ldots, \theta_{T-1}$, whereas the CCE approach involves $N(k+1)$ additional parameters, which may be a much larger number of parameters, in particular if N is large relative to T.

Equation (10) can be estimated as a linear equation by ignoring the nonlinear relationship $\boldsymbol{\delta}_t = \theta_t\boldsymbol{\beta}$ and treating $\boldsymbol{\delta}_t$ as additional parameters, cf. Hayakawa (2012). Furthermore, as the regressor $y_{i,T}$ is correlated with the errors, an instrumental variable approach is required for estimating the coefficients efficiently. Since it is assumed that $\boldsymbol{x}_{it}$ is strictly exogenous, we employ observations of all time periods to construct the $Tk \times 1$ instrumental variable vector $\boldsymbol{z}_i = (\boldsymbol{x}_{i1}', \boldsymbol{x}_{i2}', \ldots, \boldsymbol{x}_{iT}')'$. The first stage regression yields $\widehat{y}_{iT} = \widehat{\boldsymbol{\pi}}'\boldsymbol{z}_i$, where $\widehat{\boldsymbol{\pi}}'\boldsymbol{z}_i$ is the fitted value from a regression of y_{iT} on $\boldsymbol{z}_i$. The second stage regression is

$$y_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \theta_t\widehat{y}_{iT} - \theta_t\boldsymbol{\beta}'\boldsymbol{x}_{iT} + \nu_{it}.$$

Estimating the latter equation by OLS yields the two-stage least squares (2SLS) estimator. Since the error term ν_{it} is autocorrelated (due to the common component $\theta_t u_{iT}$), a GMM estimator based on the moment condition $\mathbb{E}(\boldsymbol{\nu}_i \otimes \boldsymbol{z}_i) = \mathbf{0}$ with $\boldsymbol{\nu}_i = (\nu_{i1}, \ldots, \nu_{iT})'$ is more efficient, in general.

2.4 The RS estimator

The GMM estimator of Robertson and Sarafidis (2015) results from multiplying the original model by the vector of instruments $\boldsymbol{z}_i$ (e.g. the instruments of the ALS estimator) such that

$$\begin{aligned} \boldsymbol{z}_i y_{it} &= \left(\boldsymbol{z}_i \boldsymbol{x}_{it}'\right)\boldsymbol{\beta} + \left(\boldsymbol{z}_i \lambda_i\right) f_t + \boldsymbol{z}_i u_{it} \\ \widetilde{y}_i &= \widetilde{\boldsymbol{X}}_i \boldsymbol{\beta} + \boldsymbol{\gamma}_i f_t + \widetilde{\boldsymbol{u}}_i \end{aligned} \tag{11}$$

where $\widetilde{y}_i = \boldsymbol{z}_i y_{it}$, $\widetilde{\boldsymbol{X}}_i = \boldsymbol{z}_i \boldsymbol{x}_{it}'$, $\boldsymbol{\gamma}_i = \boldsymbol{z}_i \lambda_i$, and $\widetilde{\boldsymbol{u}}_i = \boldsymbol{z}_i u_{it}$. The transformed model (11) results in a new factor model with $m = \dim(\boldsymbol{z}_i) = kT$ observations in N cross section units. In this model $\boldsymbol{\gamma}_i$ and f_t are treated as unknown parameters and the GMM estimator results from minimizing the criterion function

$$Q(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_N, f_1, \ldots, f_T) = \left(\sum_{i=1}^{N} \widetilde{y}_i - \widetilde{\boldsymbol{X}}_i \boldsymbol{\beta} - \boldsymbol{\gamma}_i f_t\right)' \boldsymbol{W}_N \left(\sum_{i=1}^{N} \widetilde{y}_i - \widetilde{\boldsymbol{X}}_i \boldsymbol{\beta} - \boldsymbol{\gamma}_i f_t\right),$$

where $\boldsymbol{W}_N$ is a consistent estimator of the optimal weighting matrix $\left[\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} \widetilde{\boldsymbol{u}}_i \widetilde{\boldsymbol{u}}_i'\right)\right]^{-1}$. Robertson and Sarafidis (2015) propose to minimize the function $Q(\cdot)$ by applying a sequential least-squares (SLS) estimator. Let f_t^0 denote some starting value. Replacing f_t by f_t^0 the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_i$ can be estimated by OLS from (11). Replacing $\boldsymbol{\gamma}_i$ by the respective OLS estimator, we can obtain an updated estimator for f_t from running T cross-section regressions (11) for $t = 1, \ldots, T$. A linear variant of this estimation approach is proposed by Juodis and Sarafidis (2020).

It is important to notice that the first order condition of the SLS estimator is invariant to some scaling factor c, such as $f_t^* = c f_t$ and $\lambda_i^* = \lambda_i / c$. The PC estimator implies $c = 1/\sqrt{\sum_{t=1}^{T} f_t^2}$ and the original ALS estimator imposes $c = 1/f_T$. The objective function of the least-squares estimator does not impose any normalization of the factors. There exists a unique minimum for the product $\boldsymbol{\gamma}_i f_t$, but the decomposition into $\boldsymbol{\gamma}_i$ and f_t is somewhat arbitrary and depends on the starting value of iterative algorithm.

3 Asymptotic properties for fixed T

The asymptotic properties of the PC and CCE estimators are typically derived by adopting a joint limit theory, where T and N tend to infinity (e.g. Pesaran 2006, Bai 2009, Greenaway-McGrevy et al. 2012 and Westerlund and Urbain 2015). The asymptotic analysis revealed that the PC and CCE estimators are $\sqrt{NT}$-consistent whenever $\sqrt{T}/N \to 0$ and $\sqrt{N}/T \to 0$. This requirement is fulfilled if for some fixed constant, $0 < a < \infty$, the paths of the sample sizes admit the inequality $aT^{0.5+\epsilon} < N < aT^{2-\epsilon}$ for some $\epsilon > 0$. Statistical inference based on these estimators suffers from an asymptotic bias whenever $T/N \to \kappa > 0$. This bias does not show up in the asymptotic analysis of Pesaran (2006), as he assumes that the coefficient vector $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{v}_i$ is individual specific, where $\boldsymbol{v}_i$ is a random error that prevents the estimator from achieving the usual $\sqrt{NT}$ convergence rate. In the literature cited above, bias-corrected estimators are suggested that remove the asymptotic bias from the limiting distribution.

For fixed T and $N \to \infty$ the CCE estimator of the factors is consistent as $\overline{e}_t(\boldsymbol{\lambda}_0)$ converges in probability to cf_t, where c is some scale factor that is different from zero. Therefore, the errors-in-variable problem vanishes for $N \to \infty$ and fixed T (cf. Westerlund et al. 2019).

For the asymptotic analysis of the PC estimator, it is usually assumed that $\min(N, T) \to \infty$ (cf. Bai 2009) and, therefore, the PC estimator may be inconsistent if T is fixed and $N \to \infty$ (see Remark 1 of Bai 2009). Under more restrictive assumptions it is however possible to show that the PC estimator of the factors is consistent if T is fixed and $N \to \infty$. To focus on the main issues assume that $\boldsymbol{\beta}$ is known. Furthermore, we assume that the vectors $\boldsymbol{f} = (f_1, \ldots, f_T)'$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$ are parameter vectors to be estimated. The PC estimator solves the first order conditions:

$$\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{e}_i - \widehat{\boldsymbol{f}}\widehat{\lambda}_i)\widehat{\lambda}_i = \boldsymbol{0} \quad \text{where } \boldsymbol{e}_i = (e_{i1}, \ldots, e_{iT})' \tag{12}$$

$$\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{e}_t - \widehat{f}_t\widehat{\boldsymbol{\lambda}})\widehat{f}_t = \boldsymbol{0} \quad \text{where } \boldsymbol{e}_t = (e_{1t}, \ldots, e_{Nt})', \tag{13}$$

subject to $T^{-1}\sum_{t=1}^{T}\widehat{f}_t^2 = T^{-1}\widehat{\boldsymbol{f}}'\widehat{\boldsymbol{f}} = 1$. Since $\widehat{\lambda}_i = T^{-1}\widehat{\boldsymbol{f}}'\boldsymbol{e}_i$, we obtain

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{e}_i - \frac{1}{T}\widehat{\boldsymbol{f}}\widehat{\boldsymbol{f}}'\boldsymbol{e}_i\right)\boldsymbol{e}_i'\boldsymbol{f} \;=\; \boldsymbol{M}_{\widehat{\boldsymbol{f}}}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{e}_i\boldsymbol{e}_i'\right)\widehat{\boldsymbol{f}} \;=\; \boldsymbol{0}, \tag{14}$$

where $\boldsymbol{M}_{\widehat{\boldsymbol{f}}} = \boldsymbol{I}_T - T^{-1}\widehat{\boldsymbol{f}}\widehat{\boldsymbol{f}}'$ with $\boldsymbol{M}_{\widehat{\boldsymbol{f}}}\widehat{\boldsymbol{f}} = \boldsymbol{0}$. For $N \to \infty$ we have

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{e}_i\boldsymbol{e}_i' \xrightarrow{p} \sigma_\lambda^2\boldsymbol{f}\boldsymbol{f}' + \boldsymbol{\Sigma}_u,$$

where $\sigma_\lambda^2 = \operatorname*{plim}_{N\to\infty} N^{-1}\sum_{i=1}^{N}\lambda_i^2$, $\boldsymbol{\Sigma}_u = \operatorname*{plim}_{N\to\infty} N^{-1}\sum_{i=1}^{N}\boldsymbol{u}_i\boldsymbol{u}_i'$, and $\boldsymbol{u}_i = (u_{i1},\ldots,u_{iT})'$. Assume that u_{it} is i.i.d. with $\boldsymbol{\Sigma}_u = \mathbb{E}(u_{it}^2)\boldsymbol{I}_T$. As $N \to \infty$ the moment condition is solved by letting $\widehat{\boldsymbol{f}} = \boldsymbol{f}$ and, therefore, the PC estimator for $\boldsymbol{f}$ is consistent (up to a scaling factor). If u_{it} is heteroskedastic or autocorrelated, then $\boldsymbol{M}_{\boldsymbol{f}}\boldsymbol{\Sigma}_u\boldsymbol{f} \neq 0$ in general and, therefore, the PC estimator is inconsistent as $N \to \infty$. On the other hand, if both N and T tend to infinity, the PC estimator is consistent no matter of a possible heteroskedasticity or (weak) autocorrelation (cf. Chamberlain and Rothschild 1983).

The asymptotic theory for the HNR and ALS estimators assumes that T is fixed and N tends to infinity. The GMM estimator is based on $kT(T-1)$ moment conditions with $k + T - 1$ unknown parameters. Therefore, no problem arises if T is fixed and N tends to infinity. Accordingly, the estimators are asymptotically normally distributed and centered around zero. Of course the problem of instrument proliferation arises if T gets large and the asymptotic theory breaks down if $T^3/N \to \kappa > 0$ (cf. Bekker 1994 and Lee et al. 2017).[5]

4 Identification

All estimation approaches require some normalization of the factors or loadings some of which may be problematical in empirical practice. The CCE and ALS

[5] A practical solution is to reduce the set of instruments (cf. Juodis and Sarafidis 2018) or applying other methods of dimensionality reduction (Breitung 2015, Section 15.2.3).

approaches are based on the following conditions:

$$\text{CCE:} \qquad \frac{1}{N}\sum_{i=1}^{N}\lambda_{0,i}\lambda_i \neq 0, \tag{15}$$

$$\text{ALS:} \qquad f_T \neq 0\ , \tag{16}$$

whereas the restriction for the PC estimator $T^{-1}\sum_{t=1}^{T} f_t^2 = 1$ is unproblematic in practice. The violation of the restrictions (15) and (16) may result in poor distributional properties of the estimator. If, for example, $N^{-1}\sum \lambda_{0,i}\lambda_i = 0$, then the cross section mean $\overline{e}_t(\boldsymbol{\lambda}_0)$ does not depend on the factor and, therefore, the CCE estimator is biased whenever $\boldsymbol{x}_{it}$ and $\lambda_i f_t$ are correlated (cf. Westerlund and Urbain 2013). Similarly, if $f_T = 0$, then $y_{iT} = \boldsymbol{\beta}'\boldsymbol{x}_{iT} + u_{iT}$ and the instruments are not able to identify the parameters θ_t and $\boldsymbol{\delta}_t$.

One may argue that the chance that (15) or (16) is exactly zero is negligible, so that problems only occur in rare cases (if at all). Unfortunately, this is not true, as the problems already arise whenever $N^{-1}\sum \lambda_{0,i}\lambda_i = O_p(N^{-1/2})$. For illustration, let us assume $\lambda_{0,i} = 1$, such that $\overline{y}_t(\boldsymbol{\lambda}_0) = \overline{y}_t$ and $\overline{\lambda} = O_p(N^{-1/2})$. Including the cross-section averages $\overline{y}_t$ and $\overline{\boldsymbol{x}}_t$ is equivalent to augmenting with $\overline{e}_t$ and $\overline{\boldsymbol{x}}_t$. Furthermore,

$$\begin{aligned}\overline{e}_t &= \overline{\lambda} f_t + \overline{u}_t \\ &= \overline{\lambda} f_t^*\ ,\end{aligned}$$

where $f_t^* = f_t + (\overline{u}_t/\overline{\lambda})$. Since in our case $\overline{u}_t/\overline{\lambda} = O(1)$, it follows that the factor f_t^* is different from f_t. In this case, $\overline{e}_t$ does not represent the true factor and the CCE estimator of $\boldsymbol{\beta}$ is inconsistent whenever the factor is correlated with the regressors.

To sidestep this difficulty, we follow the analysis of Mundlak (1978) and decompose the factor loadings into a systematic component related to the ordinary average $\overline{\boldsymbol{x}}_i$ and the projection error ξ_i:

$$\lambda_i = \gamma_0 + \boldsymbol{\gamma}_1'\overline{\boldsymbol{x}}_i + \xi_i, \tag{17}$$

where $\overline{\boldsymbol{x}}_i = T^{-1}\sum_{t=1}^{T}\boldsymbol{x}_{it}$ and ξ_i is uncorrelated with $\overline{\boldsymbol{x}}_i$. In this specification $\boldsymbol{\gamma}_1'\overline{\boldsymbol{x}}_i$ represents a possible linear dependence of λ_i on the regressors that gives rise to

an endogeneity bias. Inserting (17) in (1) yields

$$y_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \lambda_i^* f_t + e_{it}^* \ ,$$

where $\lambda_i^* = \gamma_0 + \boldsymbol{\gamma}_1'\overline{\boldsymbol{x}}_i$, $e_{it}^* = \xi_i f_t + u_{it}$ and $\mathbb{E}(e_{it}|\boldsymbol{x}_{it}) = 0$. This estimation equation is related to the projection approach of Hayakawa (2012), who considers a projection of λ_i on the vector $\boldsymbol{z}_i = vec(\boldsymbol{X}_i)$, also known as Chamberlain projection. A second difference to the Hayakawa (2012) approach is that he employs the projection for GMM estimation of ALS, whereas we employ the Mundlak projection in the context of CCE estimation.

The weighting scheme for the CCE estimator results as

$$\begin{aligned}
\overline{y}_t(\boldsymbol{\lambda}^*) &= \frac{1}{N\overline{\lambda_*^2}} \sum_{i=1}^{N} \lambda_i^* y_{it} \\
&= \widetilde{\gamma}_0 \left(\frac{1}{N} \sum_{i=1}^{N} y_{it} \right) + \widetilde{\boldsymbol{\gamma}}_1' \left(\frac{1}{N} \sum_{i=1}^{N} \overline{\boldsymbol{x}}_i y_{it} \right) \\
\text{where } \widetilde{\gamma}_0 = \gamma_0/\overline{\lambda_*^2} \ \text{ and } \ \widetilde{\boldsymbol{\gamma}}_1 &= \boldsymbol{\gamma}_1/\overline{\lambda_*^2}
\end{aligned}$$

and $\overline{\lambda_*^2} = \frac{1}{N}\sum_{i=1}^{N}(\lambda_i^*)^2$. Since $\widetilde{\gamma}_0$ and $\widetilde{\boldsymbol{\gamma}}_1$ are unknown, we augment the regression by the following $(k+1)^2$ cross section averages:

$$\begin{array}{cccc}
\frac{1}{N}\sum_{i=1}^{N} y_{it} \ , & \frac{1}{N}\sum_{i=1}^{N} x_{1,it} \ , & \cdots \ , & \frac{1}{N}\sum_{i=1}^{N} x_{k,it} \ , \\
\frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} y_{it} \ , & \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} x_{1,it} \ , & \cdots \ , & \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} x_{k,it} \ , \\
\vdots & \vdots & & \vdots \\
\frac{1}{N}\sum_{i=1}^{N} \overline{x}_{k,i} y_{it} \ , & \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{k,i} x_{1,it} \ , & \cdots \ , & \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{k,i} x_{k,it} \ .
\end{array}$$

This estimator is referred to as CCE(M).[6]

Similar normalization problems arise for the HNR and ALS approaches, but these estimators apply a normalization to the *factors*. For example, if f_T is zero,

[6]This estimator can be seen as a special case of the combination-CCE estimator proposed by Karabiyik et al. (2019).

then the linear combination of y_{iT} and $\boldsymbol{x}_{iT}$ is not able to identify the factor and, therefore, the ALS approach is biased whenever $f_T = 0$ and $\boldsymbol{x}_{it}$ is correlated with $\lambda_i f_t$. If T is small then one may try out all possible time periods for normalization and select the normalization that minimizes the GMM objective function. For a large number of time series this approach is rather time consuming. In such cases the normalization may be selected by estimating the factor by the PC approach. Then, the normalization period with the largest factor (in absolute value) is selected as the normalization period.

In the appendix of Ahn et al. (2013) a more flexible approach is proposed, which we refer to as ALS*. Let $\boldsymbol{H}$ denote the $T \times (T-1)$ orthogonal complement of $\boldsymbol{f} = (f_1, \ldots, f_T)'$ such that $\boldsymbol{H}'\boldsymbol{f} = \boldsymbol{0}$. To obtain (10) we let

$$\boldsymbol{H}'_{\text{ALS}} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -\theta_1 \\ 0 & 1 & 0 & \cdots & 0 & -\theta_2 \\ \vdots & & \ddots & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & -\theta_{T-1} \end{pmatrix}.$$

To avoid normalizing $T-1$ elements to unity, we transform the equations for unit i by using a more general matrix with property $\boldsymbol{H}'\boldsymbol{f} = \boldsymbol{0}$, such that $\boldsymbol{H}'\boldsymbol{e}_i = \boldsymbol{H}'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})'$, $\widetilde{\boldsymbol{e}}_i = \boldsymbol{H}'\boldsymbol{e}_i$. Given $\boldsymbol{\beta}$, the estimator of $\boldsymbol{H}$ is based on the moment condition $\mathbb{E}(\boldsymbol{H}'\boldsymbol{e}_i\boldsymbol{z}_i') = \boldsymbol{0}$, where $\boldsymbol{z}_i = vec(\boldsymbol{X}_i)$. Accordingly, a GMM estimator for $\boldsymbol{H}$ can be obtained as

$$\widehat{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmin}} \left\{ tr\left(\boldsymbol{H}'\boldsymbol{\Omega}_{ez}\boldsymbol{\Omega}_{zz}^{-1}\boldsymbol{\Omega}_{ez}'\boldsymbol{H}\right)\right\} \quad \text{s.t.} \quad \boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_{T-1},$$

where $\boldsymbol{\Omega}_{ez} = N^{-1}\sum_{i=1}^N \boldsymbol{e}_i\boldsymbol{z}_i'$ and $\boldsymbol{\Omega}_{zz} = N^{-1}\sum_{i=1}^N \boldsymbol{z}_i\boldsymbol{z}_i'$. Accordingly, the estimator $\widehat{\boldsymbol{H}}$ is obtained as the matrix of eigenvectors corresponding to the smallest $T-1$ eigenvalues of the matrix $\boldsymbol{\Omega}_{ez}\boldsymbol{\Omega}_{zz}^{-1}\boldsymbol{\Omega}_{ez}'$. Given $\widehat{\boldsymbol{H}}$, the estimator for $\boldsymbol{\beta}$ is obtained from the OLS regression

$$\widehat{\boldsymbol{H}}'\boldsymbol{y}_i = \widehat{\boldsymbol{H}}'\boldsymbol{X}_i\boldsymbol{\beta} + \widetilde{\boldsymbol{e}}_i.$$

This estimation step yields an updated estimator for $\boldsymbol{\beta}$ that can be used to obtain a new estimator of $\boldsymbol{H}$, until convergence. A drawback of this variant of the ALS estimator is that no standard errors for $\boldsymbol{\beta}$ are readily available, as the respective

estimation step is affected by the estimation error in $\widehat{\boldsymbol{H}}$.

It is interesting to compare this approach to the PC estimator of Bai (2009), which can be obtained by solving the problem

$$\widetilde{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmin}} \left\{ tr\left(\boldsymbol{H}'\boldsymbol{\Omega}_{ee}\boldsymbol{H}\right)\right\} \quad \text{s.t.} \quad \boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_{T-1},$$

where $\boldsymbol{\Omega}_{ee} = N^{-1}\sum_{i=1}^{N} \boldsymbol{e}_i\boldsymbol{e}_i'$. Accordingly, the difference between the PC and ALS/RS approaches is that the former extracts the factors from the residual vector $\boldsymbol{e}_i$, whereas the ALS/RS approach first projects the residuals on the space spanned by the vector of instruments $\boldsymbol{z}_i$. Accordingly, the latter approach requires that the factors are correlated with the regressors, whereas the PC approach does not.

Robertson and Sarafidis (2015) show that their estimator considered in Section 2.4 is asymptotically equivalent to ALS* if the error u_{it} is i.i.d. If u_{it} is heteroskedastic and/or serially correlated, then the weighting matrix $\boldsymbol{W}_n$ results in an asymptotic efficiency gain.

5 Multiple factors

So far we assumed that there is only a single factor. It is not difficult to see that for a panel data model with a vector of $r \geq 1$ factors $\boldsymbol{f}_t$ and the conformable $r \times 1$ loading vector $\boldsymbol{\lambda}_i$, the estimation equation (3) is given by

$$y_{it} - \boldsymbol{\lambda}_i'\overline{\boldsymbol{y}}_t^*(\boldsymbol{\Lambda}) = \boldsymbol{\beta}'\left[\boldsymbol{x}_{it} - \boldsymbol{\lambda}_i\overline{\boldsymbol{X}}_t^*(\boldsymbol{\Lambda})\right] + u_{it} - \boldsymbol{\lambda}_i\overline{u}_t(\boldsymbol{\Lambda}), \tag{18}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)'$ and

$$\overline{\boldsymbol{y}}_t^*(\boldsymbol{\Lambda}) = \left(\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'\right)^{-1}\sum_{i=1}^{N}\boldsymbol{\lambda}_i y_{it}$$
$$\text{and} \quad \overline{\boldsymbol{X}}_t^*(\boldsymbol{\Lambda}) = \left(\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'\right)^{-1}\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{x}_{it}'$$

and the $r \times 1$ vector $\overline{u}_t(\boldsymbol{\Lambda})$ is constructed in a similar manner. This shows that efficient estimation requires r linear independent weighting schemes applied to $\boldsymbol{y}_t = (x_{1t}, \ldots, y_{Nt})'$ and $\boldsymbol{X}_t = (\boldsymbol{x}_{1t}', \ldots, \boldsymbol{x}_{Nt}')'$.

To show consistency of the modified CCE estimator, CCE(M), a different reasoning is required. For the ease of exposition assume $k = 2$ regressors and $r = 2$ factors. We obtain 2 different weighting schemes:

$$\overline{y}_t^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} y_{it} \qquad \overline{x}_{1,t}^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} x_{1,it} \qquad \overline{x}_{2,t}^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} x_{2,it}$$
$$\overline{y}_t^{(2)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{2,i} y_{it} \qquad \overline{x}_{1,t}^{(2)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{2,i} x_{1,it} \qquad \overline{x}_{2,t}^{(2)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{2,i} x_{2,it}$$

that are used to obtain the following relationships:

$$\begin{pmatrix} \overline{y}_t^{(1)} \\ \overline{y}_t^{(2)} \end{pmatrix} - \begin{pmatrix} \overline{x}_{1,t}^{(1)} & \overline{x}_{2,t}^{(1)} \\ \overline{x}_{1,t}^{(2)} & \overline{x}_{2,t}^{(2)} \end{pmatrix} \beta = \begin{pmatrix} \xi_1^{(1)} & \xi_2^{(1)} \\ \xi_1^{(2)} & \xi_2^{(2)} \end{pmatrix} \begin{pmatrix} f_{1,t} \\ f_{2,t} \end{pmatrix} + O_p(N^{1/2})$$

where $\Xi_k^{(\ell)} = N^{-1}\sum_{i=1}^{N} \overline{x}_{\ell,i}\lambda_{k,i}$. Accordingly, if the matrix

$$\Xi = \begin{pmatrix} \xi_1^{(1)} & \xi_2^{(1)} \\ \xi_1^{(2)} & \xi_2^{(2)} \end{pmatrix}$$

is invertible, we can obtain the linear combinations that represent the factors as

$$\begin{pmatrix} f_{1,t} \\ f_{2,t} \end{pmatrix} = \Xi^{-1} \begin{pmatrix} \overline{y}_t^{(1)} \\ \overline{y}_t^{(2)} \end{pmatrix} - \Xi^{-1} \begin{pmatrix} \overline{x}_{1,t}^{(1)} & \overline{x}_{2,t}^{(1)} \\ \overline{x}_{1,t}^{(2)} & \overline{x}_{2,t}^{(2)} \end{pmatrix} \beta + O_p(N^{1/2})$$

Thus, the common component $\lambda_{1,i}f_{1,t} + \lambda_{2,i}f_{2,t}$ can be (asymptotically) represented by a linear combination of the 6 means $\overline{y}_t^{(1)}$, $\overline{y}_t^{(2)}$, $\overline{x}_{1,t}^{(1)}$, $\overline{x}_{2,t}^{(1)}$, $\overline{x}_{1,t}^{(2)}$, and, $\overline{x}_{2,t}^{(2)}$.[7]

6 Determining the number of factors

As argued by Pesaran (2006), the CCE estimator is consistent if the actual number of factors r is not larger than $k + 1$. This requires however that $r - 1$ factors

[7]The alert reader may have noticed that the linear combination does not involve the ordinary cross-section averages $N^{-1}\sum_i y_{it}$, $N^{-1}\sum_i x_{1,it}$ and $N^{-1}\sum_i x_{2,it}$ that are employed in the CCE estimator. These additional means are not required for identification but often improve the statistical properties of the estimator. They may also help to escape the problems resulting from a (nearly) singular matrix Ξ.

are correlated with the k regressors. This is due to the fact that one factor can be identified by the cross-section average $\overline{e}_t(\boldsymbol{\lambda}_0) = \overline{y}_t(\boldsymbol{\lambda}_0) - \boldsymbol{\beta}'\overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}_0)$, whereas the identification of the other factors requires some relationship to the cross-section averages of the regressors $\overline{\boldsymbol{x}}_t$. Furthermore, the correlation pattern needs to be sufficiently informative for identifying the factors.

It is often argued that the CCE approach is attractive, as we do not need to select the number of factors, whereas for all other approaches, the number of factors needs to be known (or determined from the data). If the number of factors is smaller than $k+1$ and the normalization requirements are satisfied, then the CCE estimator is consistent, but the small sample properties may suffer from including many cross-section averages. This is comparable to applying the PC estimator with $r = k+1$ factors. As shown by Moon and Weidner (2015), under some additional assumptions,[8] the PC estimator is robust against over-specifying the number of factors. A similar result is obtained by Westerlund et al. (2019) for the CCE estimator. Since under certain conditions the CCE estimator for $\boldsymbol{\beta}$ is as efficient as the OLS estimator using the true factors, there is no gain in (asymptotic) efficiency by changing the weighting scheme or imposing nonlinear restrictions to the auxiliary parameters that are implied by knowing the number of factors. It is however not clear whether this result provides a good guidance for empirical applications in finite samples.

In practice, it may therefore be interesting to estimate the number of factors. To this end we may invoke the criteria proposed by Bai and Ng (2002) and Ahn and Horenstein (2013). Both approaches are based on the eigenvalues of the residual covariance matrix. Denote by $\widehat{\mu}_1 \geq \cdots \geq \widehat{\mu}_T$ the ordered eigenvalues of the $T \times T$ sample covariance matrix $\widehat{\boldsymbol{\Omega}}_{ee} = N^{-1}\sum_{i=1}^{N} \widehat{\boldsymbol{e}}_i\widehat{\boldsymbol{e}}_i'$, where the residual vector $\widehat{\boldsymbol{e}}_i$ is obtained by estimating the model with maximum number of factors r^*. Furthermore, let

$$\widehat{\sigma}_u^2(r) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{u}_{it}^2 \;=\; \frac{1}{T}\sum_{j=r+1}^{T}\widehat{\mu}_j$$

where $\widehat{u}_{it}$ denotes the residual from estimating the model with r factors. Bai and

[8] The proof of Moon and Weidner (2015) requires $T \to \infty$ and is based on the i.i.d. assumption but they note that it appears that their results extend to a less restrictive setting.

Ng's (2002) criterion IC_{p2} minimizes

$$\mathrm{BN}(r) = \log\big(\widehat{\sigma}_u^2(r)\big) + r\,\frac{N+T}{NT}\log(min[N,T]),$$

for $r \in \{0, 1, \ldots, r^*\}$, whereas the criterion proposed by Ahn and Horenstein (2013) maximizes the eigenvalue ratios

$$\mathrm{AH}(r) = \widehat{\mu}_j/\widehat{\mu}_{j+1} \quad \text{for } r \in \{1, 2, \ldots, r^*\}$$

and the mock eigenvalue $\widehat{\mu}_0 = \left(\sum_{j=1}^{T}\widehat{\mu}_j\right)/\log(T)$. Let r_0 denote the true number of factors. If $\widehat{\boldsymbol{\beta}}_* - \boldsymbol{\beta} = O_p(1/\sqrt{NT})$, we have

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \widehat{\boldsymbol{\beta}}_*'\boldsymbol{x}_{it})^2 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}^2 - 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}\boldsymbol{x}_{it}'(\widehat{\boldsymbol{\beta}}_* - \boldsymbol{\beta}) + O_p\left(\frac{1}{NT}\right)\\ &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}^2 + O_p\left(\frac{1}{\sqrt{NT}}\right).\end{aligned}$$

Accordingly, the BN and AH criteria include an additional term of order $O_p((NT)^{-1/2})$ that does not affect the asymptotic properties as N and T tend to infinity.

Let us consider the asymptotic properties of the respective estimators $\widehat{r}$ if T is fixed and $N \to \infty$. In this case $\lim_{N\to\infty} P(\widehat{r} < r_0) = 0$ is ensured by (cf. Bai and Ng 2002)

$$c(N,T) = \frac{N+T}{NT}\log(min[N,T]) \to 0. \tag{19}$$

As condition (19) is not satisfied for fixed T, the BN criterion may select some $\widehat{r} < r_0$, even if $N \to \infty$. The requirement $\lim_{N\to\infty} P(\widehat{r} > r_0) = 0$ implies

$$\lim_{N\to\infty} P\left((r - r_0)c(N,T) + \log\big(\widehat{\sigma}_u^2(r)\big) - \log\big(\widehat{\sigma}_u^2(r_0)\big) > 0\right) = 1 \quad \text{for all } r > r_0. \tag{20}$$

Since $\log\big(\widehat{\sigma}_u^2(r_0)\big) - \log\big(\widehat{\sigma}_u^2(r)\big) = O_p(N^{-1}) + O_p(T^{-1})$ for $r > r_0$ (cf. Lemma 4 of Bai and Ng 2002), it may happen that for small T, condition (20) is violated as well. Hence, the BN criterion may not be consistent for fixed T. In practice it is

nevertheless possible that the BN criterion selects the number of factors consistently, if the eigenvalues $\widehat{\mu}_1, \ldots, \widehat{\mu}_{r_0-1}$ are sufficiently large and $\widehat{\mu}_{r_0+1}, \ldots, \widehat{\mu}_{r^*}$ are sufficiently small relative to $\widehat{\mu}_{r_0}$.

Since for fixed T, $\widehat{\mu}_r$ is $O_p(1)$ for all $r = 1, \ldots, T$, it follows that the eigenvalue ratio AH(r) is $O_p(1)$ for fixed T and all $r \in \{1, \ldots, r^*\}$. Therefore, the AH criterion cannot be shown to be a consistent selection rule for fixed T. It may nevertheless perform well, if the slope of the eigenvalue function is sufficiently steep at $r = r_0$.

A possibility to sidestep these problems is to adopt the BIC selection criteria of Ahn et al. (2013) and Robertson and Sarafidis (2015). These criteria are based on the Sargan-Hansen specification test for GMM estimators. If the number of factors is too small, then the remaining cross-correlation among the residuals results in a large value of the test statistic. The penalty function is constructed such that the sum of the test statistic and the penalty function obtains a minimum at the correct number of factors as N tends to infinity.

7 Monte Carlo Simulations

In this section we assess the performance of alternative estimation methods in various settings and highlight some favorable and problematic aspects of alternative estimation methods. The simulation results in Sections 7.1 – 7.2 are based on the following simple data-generating process

$$y_{it} = \beta x_{it} + \lambda_i f_t + u_{it} \tag{21}$$

$$x_{it} = \mu + \lambda_i f_t + \lambda_i + f_t + \varepsilon_{it} \tag{22}$$

with $\beta = 0.5$ and $r = 1$. Hence, the regressor is correlated with the loadings, the factor and the product of both. The regression error u_{it} and the idiosyncratic component of the regressor, ε_{it}, are independent standard normal random variables. The constant μ is drawn from a $U[0, 1]$ distribution. The DGPs in Sections 7.1 to 7.2 differ with respect to the distributional assumptions on the factors and their loadings.

The (near) violation of the normalization restrictions for the CCE and ALS estimators are examined in Section 7.1. In Section 7.2, we compare the PC and

CCE estimator with regard to their different weighting schemes. In Section 7.3 we address the estimation of the number of factors, r, for the PC, ALS* and RS approaches. There, we consider a similar DGP as in (21) and (22) for $r = 1$ and $r = 2$. The last subsection 7.4 considers the relative performance of the CCE, PC, ALS* and RS, estimation approaches in more general settings that are based on the DGPs considered by Bai (2009), Chudik et al. (2011) and Ahn et al. (2013).

7.1 Normalization failure

As argued in Section 4, the CCE and ALS/HRN approaches may suffer from a violation of their normalization conditions. The performance already deteriorates if the parameters approach the $\sqrt{N}$-vicinity of the problematic subspace. In a model with a single factor, the normalization of the equally weighted CCE estimator ($\lambda_{0,i} = 1$) requires that $\overline{\lambda} = N^{-1}\sum_{i=1}^{N}\lambda_i \neq 0$. We have argued that whenever $\overline{\lambda} = c/\sqrt{N}$, the factor cannot be represented by a linear combination of $\overline{y}_i$ and $\overline{x}_i$ as $N \to \infty$.

Sarafidis and Wansbeek (2012) and Westerlund and Urbain (2013) analyze the performance of the CCE estimator when the normalization condition is violated. In order to study the performance of the CCE estimator when $\overline{\lambda}$ is different but close to zero, we consider the model in (21) and (22), where we generate the factor loadings as

$$\text{DGP1:} \quad \lambda_i \sim \mathcal{N}(\mu_\lambda, 1) \text{ for } \mu_\lambda \in [0,1] \quad \text{and} \quad f_t \sim \mathcal{N}(0,1).$$

Hence, the loadings are normally distributed with expectation that ranges from 0 to 1.

Figures 1 (a) – (d) present the absolute bias for the original CCE, the Mundlak type CCE(M) estimator suggested in Section 4, and the PC estimator for $N = 100$ and $N = 500$ with a small ($T = 10$) and moderate ($T = 50$) number of time periods. The PC estimator of Bai (2009) is obtained by a sequential estimation procedure using the pooled OLS estimator as starting value for β (see Section 2.1). It turns out that the CCE estimator is severely biased even if the mean of λ_i is substantially different from zero. This is due to the fact that a bias already occurs whenever $\mu_\lambda = O(N^{-1/2})$. This reasoning predicts that for fixed μ_λ the bias gets smaller if N increases. Indeed, this is what we observe when comparing

panel (a) and (c) as well as (b) and (d). Note that $\sqrt{100}/\sqrt{500} \approx 0.44$ and, therefore, we expect that the bias reduces to a value less than one half which is a good approximation for $\mu_\lambda > 0.1$. The other two estimators, PC and CCE(M), are virtually unbiased, which is expected as the estimators do not rely on the assumption $\mu_\lambda \neq 0$.

In a similar manner, the normalization of the ALS estimator may be problematic if the factors approach the problematic subspace. The ALS estimator requires $f_T \neq 0$. To examine the consequences of an (approximate) violation of this normalization condition, we consider the model in (21) and (22) where the factors are generated as:

$$\text{DGP2: } f_t \sim \mathcal{N}(0,1) \text{ for } t = 1, ..., T-1 \text{ and } f_T \sim \mathcal{N}(\mu_T, 0.5) \text{ for } \mu_T \in [0,1]$$

and the factor loadings are standard normally distributed. As the final value of the factor is crucial, we generate it by a distribution with expectation ranging from 0 to 1.

Figures 1 (e) – (f) present the bias for the ALS estimator when $T = 5$ and $N = 100$ or $N = 500$, respectively. As expected, the ALS estimator is severely biased whenever $\mu_T = \mathbb{E}(f_T)$ is small. But even for moderate values of μ_T the bias remains substantial and decreases only gradually for larger values of μ_T. It should be noted that if the regression includes an individual specific intercept, then the factors are demeaned and, therefore, assuming a nonzero mean appears inappropriate.

Figures 1 (e) – (f) also present the bias of two estimators that circumvent the problems with the normalization of the original ALS estimator. The estimator ALS* refers to the GMM estimator that estimates the matrix H that is used to remove the factors (see Section 4).[9] Our simulation results suggest that this estimator performs quite well in terms of bias, as it is virtually unbiased for all values of μ_T. Another approach to escape the normalization problem is the GMM$_{max}$ estimator, where in a first step the factor is estimated using the PC approach. In the second step, the time period for the normalization is chosen according to the maximum absolute value of the estimated factor and the original ALS estimator is adapted, where the time period with the largest factor is shifted to the end of the sample. Both estimators are able to reduce the bias dramatically.

[9]Following Ahn et al. (2013), we use $\beta = 0$ as starting value for the iterative ALS* procedure.

The figures also include the RS estimator, which corresponds to the FIVU estimator of Robertson and Sarafidis (2015). This estimator does not require $f_T \neq 0$ for normalization (see Section 2.4) and thus the bias does not depend on the value of μ_T. The RS estimator has a slight advantage in terms of bias when $N = 100$. With $N = 500$, the bias of the ALS*, GMM$_{max}$ and RS estimators is nearly zero.

To summarize, our findings confirm earlier evidence that the normalization applied for the original CCE or ALS/HNR estimators may be problematical, whenever the factors or loadings approach a normalization failure. It is however easy to adjust the estimators such that they perform well for all values of the parameter space. Our Monte Carlo exercise indicates that the PC and CCE(M) estimators as well as ALS*, GMM$_{max}$ and RS are very robust against a possible normalization failure.

7.2 Fixed versus data driven weights

From the reasoning of Section 2, it turns out that the CCE estimator is expected to outperform the PC estimator whenever the weighting scheme $\boldsymbol{\lambda}_0$ comes close to the actual set of loadings $\boldsymbol{\lambda}$, see also Westerlund and Urbain (2015). For equal weights with $\lambda_{0,i} = 1$ for all i, the CCE estimator performs well, whenever (i) the absolute value of the mean of the loadings is large (to avoid the normalization failure) and (ii) the variance of the loadings is small. Our DGP3 represents such a scenario, whereas the DGP4 favors the PC estimator by generating factor loadings with large variance,

$$\begin{aligned} &\text{DGP3: } \lambda_i \sim \mathcal{N}(1, 0.1), \;\; f_t \sim \mathcal{N}(0, 1) \\ &\text{DGP4: } \lambda_i \sim \mathcal{N}(1, 3), \;\; f_t \sim \mathcal{N}(0, 1). \end{aligned}$$

The remaining details of the simulation setup are identical to the model in (21) and (22).

The results reported in Table 1 clearly confirm our assertion that the CCE estimator outperforms the PC estimator in DGP3, whereas the PC estimator performs better for DGP4. This finding suggests to find a weighting scheme that comes close to the actual distribution of the loadings. This is the notion behind the Mundlak type CCE variant that employs the individual specific means $\overline{y}_i$

and $\overline{x}_i$, since a linear combination of these averages can be seen as (CCE type) estimates of the loadings λ_i. Therefore, we hope to improve the original CCE estimator by applying weights that are correlated with the loadings. Our results from the simple Monte Carlo experiment suggest that the CCE(M) approach of choosing a data driven weighting scheme performs similar to the best estimator in the respective situation. Furthermore, as shown in the previous subsection, the CCE(M) estimator sidesteps the risk of a normalization failure. Provided that this estimator is similarly easy to compute as the original CCE estimator, it appears as if this estimator is a robust and efficient variant of the original CCE estimator.

7.3 Selecting the number of factors

In practice, it is necessary to select the number of factors for the PC and GMM estimation procedures. The choice is important, since misspecifying the number of factors can have severe consequences: Overspecifying the number of factors can have adverse effects on the sampling properties of the estimators, while an underspecification may lead to inconsistent estimates if the ignored factors are correlated with the regressors. One possibility for selecting the number of factors is simply to specify the number according to some ad hoc rule, for instance $r = k + 1$, as usually advocated for the CCE approach. Another option is to use a consistent criterion for the number of factors, such as the ones proposed by Bai and Ng (2002) (hereafter: BN) and Ahn and Horenstein (2013) (AH). Note that these selection criteria were developed for the pure factor model without regressors. Furthermore, the asymptotic theory underlying these approaches requires $T \to \infty$ (see Section 6). It is therefore interesting to investigate the performance of these criteria that were not initially developed for a small number of time periods. For the GMM estimators, the number of factors can be estimated using model information criteria, such as the Schwarz Criterion (BIC) considered by Ahn et al. (2013) and Robertson and Sarafidis (2015).

In order to study the performance of these selection criteria, we consider a similar model as in (21) and (22) with $r = 1$ and $r = 2$. For the loadings and factors, we assume the following DGP,

$$\text{DGP5: } \lambda_{j,i} \sim \mathcal{N}(0,1),\ f_{j,t} \sim \mathcal{N}(0,1) \ \text{ for } \ j = 1, 2.$$

As reported in Table 2, the hit rates for a single factor, $r = 1$, are nearly 100% for the BN and AH criteria whenever $T \geq 10$. For $T = 5$ the BN criterium does not work and nearly always picks the maximum number of factors. On the other hand the AH criterion works remarkably well, even for a number of time periods as small as $T = 5$.[10] The hit rates for the BIC criteria exceed 90% in all but one case. For $r = 2$ the hit rates for the AH criterion are substantially lower, but the estimators are still quite accurate, even if $T = 10$ and N is large. For the BIC criteria, the hit rates decrease by only a small amount and do not seem to be very sensitive to the number of factors, in particular if $N > 100$.

In Table 3, we report bias and RMSE for the PC, ALS* and RS estimators based on the true number of factors ($r = 1$ and $r = 2$) as a benchmark. In addition we assess the performance of the estimators, when the number of factors is estimated based on selection criteria.[11] As expected, using the AH method for $r = 1$ in order to estimate the number of factors for the PC estimator produces bias and RMSE results that are of similar magnitude as the true number of factors. Applying the BIC criterion to estimate the number of factors for the GMM estimators produces very accurate estimates when $N > 100$, accordingly.

For $r = 2$, the performance of the PC estimator using the AH criterion shows a considerable bias, in particular if T is as small as 5. In contrast, bias and RMSE of the GMM estimators applying the BIC criterion are similar to the estimators based on the true number of factors when $N > 100$. When T increases to 10, there is still a substantial performance gap between the PC estimator using the AH method and the PC estimator based on the true number of factors, whereas the GMM estimators based on the BIC criterion perform much better. This is surprising as Table 2 suggests that the hit rates of the BIC criterion are only slightly better in these cases. The reason is that the AH criterion tends to underestimate the number of factors whereas the BIC criterion overestimates the number of factors in case the correct number of factors is not found.

Consider, for instance, $T = 10$ and $N = 500$. The BIC estimator finds the

[10] The performance is similar to the case where β is known (not shown). Therefore, the estimation of β does not seem to have an important effect on the performance of the BN and AH selection criteria. Furthermore, the growth ratio statistic of Ahn and Horenstein (2013) performs similar to the eigenvalue ratio statistic. For reasons of space we do not show the respective results.

[11] To save space, we do not show results for the estimators based on the BN criterion, since the hit rates are either 0% or (close to) 100%.

correct number of factors ($r = 2$) in more than 95% of the cases and overestimates the number in the other ($< 5\%$) cases. The AH estimator finds the correct value of $r = 2$ in 89.8% of the cases, however underestimates the number in all other cases. Since the estimator is biased if the number of factors is too small, the AH criterion tends to produce a large negative bias in some cases, whereas the BIC criterion tends to produce unbiased estimators with a slightly larger variance than estimating with the correct number of factors in some very rare cases.

7.4 Performance in more general setups

So far the DGPs considered in this paper were simplified versions of the ones considered in the literature and focus on the particular features of these models. In the following, we study the relative performance of the CCE, PC, ALS* and RS approaches in more sophisticated simulation setups, similar to the simulation experiments of Bai (2009), Chudik et al. (2011) and Ahn et al. (2013). The details of these data generating processes are presented in the online appendix to this paper. The Monte Carlo design of Bai (2009) employs two regressors that are correlated with two factors, their loadings and the product of both. The idiosyncratic error is i.i.d. across individuals and time periods. We refer to this model as DGP6. DGP7 refers to the factor model of Chudik et al. (2011) that includes two regressors and three factors. A special feature of this DGP is that the factor loadings of the regressors are independent of the loadings in the errors e_{it}. Accordingly, no endogeneity bias arises from estimating the model by a pooled OLS estimator. The factors are generated by independent AR(1) processes and the idiosyncratic component u_{it} is heteroskedastic but mutually and serially uncorrelated. DGP8 corresponds to the Monte Carlo design of Ahn et al. (2013), which includes two regressors and two factors. The first regressor is correlated with the first factor and the second regressor is correlated with the second factor. The idiosyncratic error is autocorrelated but the variances are identical across panel units and time periods.

The results in Table 4 indicate that the relative performance of the estimators depends quite sensitively on the DGP considered. The first panel of Table 4 presents the results for DGP6. The CCE estimator is not consistent in this setting, since the rank condition is violated and both factor and loading vectors are correlated with both regressors. The other three estimators are consistent in

this setting, where the RS estimator is the least biased when $T = 5$ and the ALS* exhibits the lowest bias for $T \geq 10$. The latter performs best in terms of RMSE with only slight advantages over the PC estimator when $T \geq 10$.

The second panel of Table 4 reports the results for DGP7. The CCE estimator is the favored one in this setting. It has a very small bias and exhibits the lowest RMSE for nearly all considered (N, T) combinations, in particular if T is as small as 5. Comparing the PC and GMM estimators, the results slightly favor the PC estimator in terms of RMSE. The difference between the PC and the CCE estimator is negligible when $T = 15$ and $N = 500$. With regard to the GMM estimators, the RS estimator has a marginally lower RMSE when $T = 5$ and N is large, while the results indicate small advantages for the ALS* estimator when $T \geq 10$.

The third panel of Table 4 presents the results for DGP8. The GMM estimators are the least biased estimators in this setting. The ALS* estimator exhibits the smallest RMSE for all (N, T) combinations with only slight advantages over the RS estimator. For example, for $T = 10$ and $N = 500$, the RMSE of the ALS* estimator is about 40% lower than the RMSE of the PC estimator and more than 60% lower than the RMSE of the CCE estimator. The CCE estimator is problematic in this setting, since the expectation of the loadings is equal to zero. The PC estimator is problematic in this small T setting. However, the RMSE is lower for larger samples with $T = 15$ and $N = 500$.

8 Conclusion

In this paper we compare three existing approaches for estimating factor augmented panel data models. We argue that the PC estimator can be seen as an estimated analog of the optimal transformation for eliminating the common factors from the data. The CCE estimator applies a data transformation that has the important advantage that the weighting scheme is fixed and does not involve any sampling error. This ensures that the estimator is consistent even if T is fixed, whereas the PC estimator requires much more restrictive assumptions (such as i.i.d. errors) if T is fixed. The third estimation approach is the nonlinear GMM estimators of Ahn et al. (2013) and Robertson and Sarafidis (2015). In contrast to the PC and CCE estimators, this estimator treats the T observations of the

factor as parameters, whereas the factor loadings are substituted out. Therefore the number of parameters involved by this approach does not depend on N.

An important difference between the PC estimator and all other approaches is that the PC estimator does not require that the factors are correlated with the regressors. In contrast, the ALS/RS approaches and the CCE estimator (for $r > 1$) rely on the assumption that the factors are linearly related to the regressors (that is, the instruments are relevant). Accordingly, if some factors are uncorrelated (or weakly correlated) with the regressors, one can expect the PC estimator to be more efficient.

In this paper we focus on the typical micro panel data setup where T is small compared to N. Since for an approximate factor model the consistency of the PC estimator requires $T \to \infty$, it is interesting to investigate how large T needs to be for ensuring the PC estimator to be approximately unbiased. Our Monte Carlo experiments indicate that for all data generating mechanisms considered in this paper $T = 10$ is already sufficient to achieve reasonable small sample properties of the PC estimator. Sometimes the CCE and ALS* estimators perform slightly better than the PC estimator, but in other Monte Carlo setups the PC estimator clearly outperforms all other competitors. Furthermore, we show that for small T the selection criteria for the number of factors proposed by Bai and Ng (2002) and Ahn and Horenstein (2013) may be inconsistent, whereas the BIC criteria of Ahn et al. (2013) and Robertson and Sarafidis (2015) perform well.

Compliance with Ethical Standards

Conflict of Interest: The author Jörg Breitung declares that he has no conflict of interest. Author Philipp Hansen declares that he has no conflict of interest.
Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ahn, S. and Horenstein, A. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

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

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

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Bai, J. (2009). Panel data model with interative fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Baltagi, B. (2005). *Econometric Analysis of Panel Data*. John Wiley & Sons Inc., New York, 3 edition.

Bekker, P. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, 62(3):657–681.

Breitung, J. (2015). The analysis of macroeconomic panel data. In Baltagi, B., editor, *The Oxford Handbook of Panel Data*, chapter 15, pages 453–492. Oxford University Press.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

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

Everaert, G. and De Groote, T. (2016). Common correlated effects estimation of dynamic panels with cross-sectional dependence. *Econometric Reviews*, 35(3):428–463.

Greenaway-McGrevy, R., Han, C., and Sul, D. (2012). Asymptotic distribution of factor augmented estimators for panel regression. *Journal of Econometrics*, 169(1):48–53.

Hayakawa, K. (2012). GMM estimation of short dynamic panel data models with interactive fixed effects. *Journal of the Japan Statistical Society*, 42(2):109–123.

Holtz-Eakin, D., Newey, W., and Rosen, H. (1988). Estimating vector autoregressions with panel data. *Econometrica*, 56(6):1371–1395.

Juodis, A. and Sarafidis, V. (2018). Fixed t dynamic panel data estimators with multifactor errors. *Econometric Reviews*, 37(8):893–929.

Juodis, A. and Sarafidis, V. (2020). A linear estimator for factor-augmented fixed-t panels with endogenous regressors. *fortcoming in: Journal of Business & Economic Statistics*.

Karabiyik, H., Urbain, J.-P., and Westerlund, J. (2019). CCE estimation of factor-augmented regression models with more factors than observables. *Journal of Applied Econometrics*, 34(2):268–284.

Lee, N., Moon, H., and Zhou, Q. (2017). Many IVs estimation of dynamic panel regression models with measurement error. *Journal of Econometrics*, 200(2):251–259.

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

Moon, H. and Weidner, W. (2019). Nuclear norm regularized estimation of panel regression models. *cemmap Working Paper*, CWP14/19.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

Pesaran, M. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Robertson, D. and Sarafidis, V. (2015). IV estimation of panels with factor residuals. *Journal of Econometrics*, 185(2):526–541.

Sarafidis, V. and Wansbeek, T. (2012). Cross-sectional dependence in panel data analysis. *Econometric Reviews*, 31(5):483–531.

Su, L. and Jin, S. (2012). Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics*, 169(1):34–47.

Westerlund, J., Petrova, Y., and Norkute, M. (2019). CCE in fixed-T panels. *Journal of Applied Econometrics*, 34(5):746–761.

Westerlund, J. and Urbain, J.-P. (2013). On the estimation and inference in factor-augmented panel regressions with correlated loadings. *Economics Letters*, 119(3):247–250.

Westerlund, J. and Urbain, J.-P. (2015). Cross-sectional averages versus principal components. *Journal of Econometrics*, 185(2):372–377.

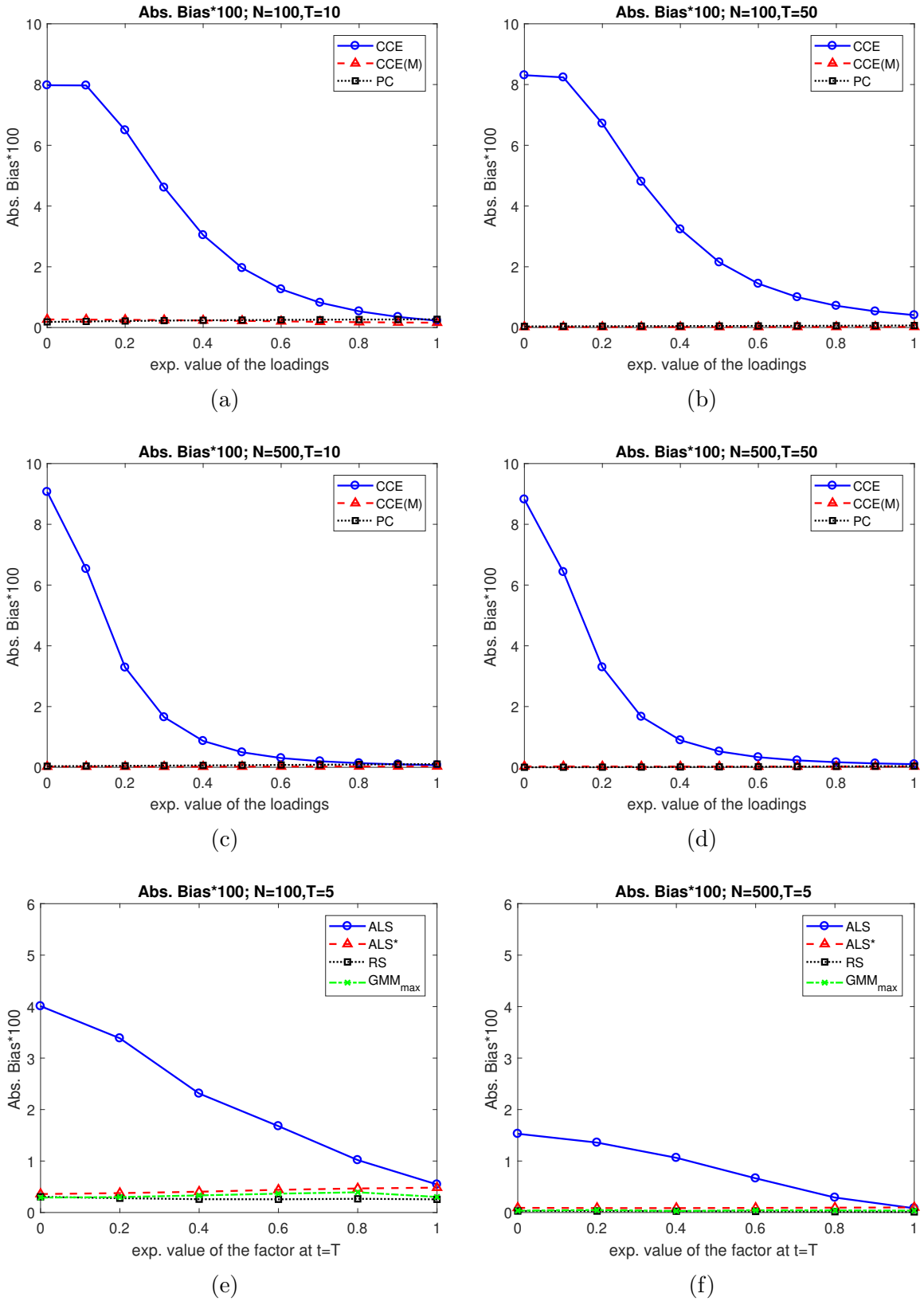

Figure 1: Normalization failure for CCE (DGP1) and ALS (DGP2)

Table 1: Fixed versus data driven weights

		Bias*100			RMSE*100		
N	T	PC	CCE	CCE(M)	PC	CCE	CCE(M)
				DGP3			
50	10	1.23	0.00	0.19	6.43	5.12	5.93
100	10	0.56	0.06	0.21	3.94	3.56	4.04
100	20	0.10	−0.14	−0.09	2.43	2.33	2.42
100	50	0.09	−0.04	0.02	1.49	1.48	1.51
100	100	0.08	−0.03	0.02	1.06	1.06	1.08
500	500	0.05	−0.01	−0.01	0.20	0.20	0.20
				DGP4			
50	10	0.18	−2.31	0.19	4.65	6.62	5.97
100	10	0.24	−1.09	0.22	3.26	4.05	4.17
100	20	0.01	−1.30	−0.08	2.15	3.05	2.45
100	50	0.08	−1.22	0.01	1.34	2.36	1.51
100	100	0.10	−1.20	0.01	0.97	2.00	1.08
500	500	0.08	−0.24	−0.01	0.20	0.36	0.20

This table reports the simulation results generated with DGPs 3 and 4. The results are based on 1000 replications.

Table 2: Hit rates for selection criteria

		r=1				r=2			
N	T	BN_{PC}	AH_{PC}	BIC_{ALS^*}	BIC_{RS}	BN_{PC}	AH_{PC}	BIC_{ALS^*}	BIC_{RS}
100	5	0.0	94.6	91.7	83.0	0.0	46.8	86.4	76.6
250	5	0.0	96.2	96.9	96.7	0.0	50.8	93.2	89.5
500	5	0.0	96.9	98.8	98.3	0.0	52.1	96.3	94.1
250	10	100.0	99.9	90.6	97.0	99.6	86.4	89.7	92.9
500	10	99.9	99.9	96.7	98.4	99.4	89.8	95.9	96.9
500	15	100.0	100.0	92.3	99.6	100.0	97.9	92.9	98.8

Table 3: Selecting the number of factors

		$r=1$				$r=2$			
		Bias*100		RMSE*100		Bias*100		RMSE*100	
N	T	PC_r	PC_{AH}	PC_r	PC_{AH}	PC_r	PC_{AH}	PC_r	PC_{AH}
100	5	0.20	0.31	5.10	5.30	0.67	5.47	6.83	11.06
250	5	0.12	0.24	3.22	3.51	0.29	4.89	4.14	10.13
500	5	0.14	0.30	2.25	2.66	0.22	4.54	3.04	9.13
250	10	0.07	0.08	2.04	2.04	0.17	1.51	2.20	4.79
500	10	0.09	0.10	1.42	1.44	0.06	1.07	1.55	3.93
500	15	0.07	0.07	1.06	1.06	0.11	0.28	1.18	1.81
		ALS^*_r	ALS^*_{BIC}	ALS^*_r	ALS^*_{BIC}	ALS^*_r	ALS^*_{BIC}	ALS^*_r	ALS^*_{BIC}
100	5	0.14	0.15	6.22	6.84	−0.33	−0.65	7.46	8.15
250	5	−0.01	0.04	3.69	3.83	0.08	0.10	4.33	4.53
500	5	0.23	0.23	2.62	2.64	0.04	−0.01	3.08	3.26
250	10	−0.02	−0.02	2.25	2.36	0.00	−0.02	2.23	2.32
500	10	0.10	0.10	1.59	1.61	−0.06	−0.06	1.58	1.59
500	15	0.04	0.03	1.20	1.22	0.03	0.03	1.18	1.20
		RS_r	RS_{BIC}	RS_r	RS_{BIC}	RS_r	RS_{BIC}	RS_r	RS_{BIC}
100	5	−0.58	0.74	6.01	7.93	−0.92	−0.26	7.88	9.00
250	5	−0.17	−0.12	3.65	3.76	−0.21	−0.14	4.72	4.99
500	5	0.11	0.10	2.60	2.66	−0.07	−0.10	3.58	3.59
250	10	−0.40	−0.29	2.42	2.68	−0.86	−0.73	3.20	3.21
500	10	−0.10	−0.10	1.66	1.66	−0.44	−0.41	2.16	2.11
500	15	−0.17	−0.17	1.29	1.29	−0.65	−0.62	2.05	2.00

This table reports bias and RMSE results for DGP5 with $r = 1$ and $r = 2$ for the PC, ALS* and RS estimators with the true number of factors and estimated number of factors based on selection criteria. The results are based on 1000 replications.

Table 4: Performance in more general setups

		Bias*100								RMSE*100							
		CCE		PC		ALS*		RS		CCE		PC		ALS*		RS	
N	T	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
DGP Bai (2009)																	
100	5	10.27	10.27	3.54	3.90	−2.14	−1.92	−0.75	−0.07	24.40	24.19	13.65	13.94	13.34	13.95	15.71	16.21
250	5	10.71	10.45	1.61	1.27	−0.69	−1.51	−0.61	−1.21	21.98	22.02	8.77	8.34	7.74	8.10	9.23	9.37
500	5	10.78	11.51	0.36	1.01	−0.69	−0.07	−0.63	−0.04	21.31	22.02	6.02	6.17	5.23	5.35	6.23	6.39
250	10	13.43	13.98	0.35	0.69	−0.21	0.14	−1.43	−1.02	17.78	18.24	4.27	4.20	4.17	4.03	5.28	4.97
500	10	13.08	13.14	0.19	0.26	−0.13	−0.05	−0.58	−0.50	17.36	17.54	3.05	2.97	2.74	2.67	3.42	3.23
500	15	13.75	13.72	0.17	0.14	−0.03	−0.05	−0.65	−0.72	16.54	16.50	2.18	2.13	2.14	2.12	2.54	2.56
DGP Chudik et al. (2011)																	
100	5	−0.28	−0.42	1.71	0.94	−1.48	−2.01	0.70	0.18	8.60	8.89	10.82	10.21	11.15	11.47	12.04	11.77
250	5	−0.04	−0.10	0.37	0.51	−1.07	−0.99	−0.01	−0.14	5.36	5.19	5.73	5.80	6.83	6.49	6.67	6.64
500	5	−0.01	−0.04	0.14	0.24	−0.22	−0.07	0.12	0.20	3.69	3.71	4.18	4.08	5.08	4.71	4.95	4.44
250	10	−0.01	−0.10	0.10	0.01	−0.07	−0.06	0.03	−0.08	2.73	2.72	2.80	2.74	3.17	3.11	3.97	3.99
500	10	0.15	0.03	0.17	0.08	0.11	0.09	0.00	−0.02	1.93	1.90	2.06	1.92	2.24	2.14	2.60	2.68
500	15	0.07	−0.04	0.13	0.04	0.03	−0.03	0.18	−0.05	1.50	1.44	1.49	1.40	1.59	1.53	2.23	2.12
DGP Ahn et al. (2013)																	
100	5	3.68	3.75	0.11	0.56	−0.23	0.08	0.00	0.06	10.50	10.98	9.54	9.66	5.73	5.78	6.75	6.42
250	5	1.63	1.84	−0.05	0.02	0.06	0.09	−0.01	0.07	6.93	6.89	6.81	7.12	3.41	3.25	3.59	3.43
500	5	0.75	0.85	0.14	−0.50	0.12	−0.03	0.13	−0.06	4.89	5.03	6.26	6.08	2.27	2.36	2.34	2.41
250	10	2.14	1.88	−0.59	−0.79	−0.03	−0.07	−0.11	−0.19	5.81	5.77	2.99	2.90	2.06	2.05	2.49	2.51
500	10	1.10	0.90	−0.67	−0.69	0.00	−0.10	0.00	−0.09	4.04	4.01	2.28	2.49	1.42	1.46	1.61	1.70
500	15	1.13	0.89	−0.47	−0.52	0.05	0.02	−0.05	−0.07	4.02	3.75	1.56	1.67	1.09	1.09	1.50	1.55

This table reports bias and RMSE results for the CCE, PC, ALS* and RS estimators generated by the DGPs 6-8. The results are based on 1000 replications.

Supplemental Material

Details of the data generating processes used in Section 7.4

DGP6, (Bai, 2009)

We consider the following model with two regressors, $k = 2$, and $r = 2$ unobserved factors:

$$y_{it} = \beta_1 x_{1,it} + \beta_2 x_{2,it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{it}, \tag{23}$$

with $\beta_1 = 1$, $\beta_2 = 3$, $\boldsymbol{\lambda}_i = (\lambda_{1,i}, \lambda_{2,i})'$ and $\boldsymbol{f}_t = (f_{1,t}, f_{2,t})'$. The two regressors are generated as

$$x_{1,it} = \mu_1 + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \boldsymbol{\iota}' \boldsymbol{\lambda}_i + \boldsymbol{\iota}' \boldsymbol{f}_t + \eta_{1,it} \tag{24}$$

$$x_{2,it} = \mu_2 + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \boldsymbol{\iota}' \boldsymbol{\lambda}_i + \boldsymbol{\iota}' \boldsymbol{f}_t + \eta_{2,it} \tag{25}$$

with $\boldsymbol{\iota}' = (1, 1)$. Hence, both regressors are correlated with the loadings, the factors and the product of both. The unobserved factors and loadings follow standard normal distributions,

$$f_{j,t} \overset{iid}{\sim} \mathcal{N}(0, 1) \text{ for } j = 1, 2,$$
$$\lambda_{j,i} \overset{iid}{\sim} \mathcal{N}(0, 1) \text{ for } j = 1, 2,$$

where $j = 1, 2$ denotes the factor subscript. The regression error is generated as

$$u_{it} \overset{iid}{\sim} \mathcal{N}(0, 4)$$

and the idiosyncratic components of the regressors are generated as

$$\eta_{l,it} \overset{iid}{\sim} \mathcal{N}(0, 1) \text{ for } l = 1, 2,$$

where l indicates the regressor and $\mu_l = 1$ for $l = 1, 2$.

DGP7, (Chudik et al., 2011)

This simulation setup is based on a model with two regressors and three unobserved factors,

$$y_{it} = \beta_1 x_{1,it} + \beta_2 x_{2,it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{it} \tag{26}$$

where $\beta_1 = \beta_2 = 1$, $\boldsymbol{\lambda}_i = (\lambda_{1,i}, \lambda_{2,i}, \lambda_{3,i})'$ and $\boldsymbol{f}_t = (f_{1,t}, f_{2,t}, f_{3,t})'$. The regressors are generated according to

$$x_{1,it} = \boldsymbol{\gamma}_{1,i}' f_t + \eta_{1,it}, \tag{27}$$

$$x_{2,it} = \boldsymbol{\gamma}_{2,i}' f_t + \eta_{2,it}, \tag{28}$$

where $\boldsymbol{\gamma}_{1,i}$ and $\boldsymbol{\gamma}_{2,i}$ denote r-dimensional vectors of loadings for the regressors that are independent of the loadings in the DGP of the dependent variable, $\boldsymbol{\lambda}_i$. The unobserved factors are generated as independent AR(1) processes,

$$f_{j,t} = 0.5 f_{j,t-1} + \upsilon_{f_{j,t}},\ j = 1, 2, 3;\ t = -49, ..., 0, 1, ..., T$$
$$\upsilon_{f_{j,t}} \overset{iid}{\sim} \mathcal{N}(0, 1 - 0.5^2),\ f_{j,-50} = 0.$$

In order to reduce the effect of the initial value, the first 50 observations of $f_{j,t}$ are discarded. The factor loadings in the DGP of y_{it} are generated as

$$\lambda_{j,i} \overset{iid}{\sim} \mathcal{N}(0, 1) \text{ for } j = 1, 2, 3$$

and are independently distributed from the factor loadings in the DGPs of the regressors,

$$\gamma_{l,j,i} \overset{iid}{\sim} \mathcal{N}(0, 1) \text{ for } l = 1, 2;\ j = 1, 2, 3$$

where l denotes the index for the regressor $x_{l,it}$. The regression errors exhibit mild heteroskedasticity and are generated as

$$u_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_i^2),\ \text{where } \sigma_i^2 \overset{iid}{\sim} \mathcal{U}(0.5, 1.5).$$

The idiosyncratic components of the regressors are generated according to

$$\eta_{l,it} = \rho_{\nu_{l,i}}\eta_{l,it-1} + \nu_{j,it} \text{ for } l = 1, 2;\ t = -49, ..., 0, 1, ...T$$
$$\nu_{l,it} \overset{iid}{\sim} \mathcal{N}(0, 1 - \rho^2_{\nu_{j,i}}),\ \eta_{l,i,-50} = 0,\ \rho_{\nu_{l,i}} \overset{iid}{\sim} \mathcal{U}(0.05, 0.95) \text{ for } l = 1, 2.$$

The first 50 observations of $\eta_{l,t}$ are discarded as "burn-in" period.

DGP8, (Ahn et al, 2013):

For this DGP, we consider a model with $k = 2$ and $r = 2$,

$$y_{it} = \beta_1 x_{1,it} + \beta_2 x_{2,it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{it} \tag{29}$$

where $\beta_1 = \beta_2 = 1$, $\boldsymbol{\lambda}_i = (\lambda_{1,i}, \lambda_{2,i})'$ and $\boldsymbol{f}_t = (f_{1,t}, f_{2,t})$. The regressors are generated by

$$x_{1,it} = \lambda_{1,i} f_{1,t} + \lambda_{1,i} + f_{1,t} + \eta_{1,it} + \mu_{1,i} \tag{30}$$
$$x_{2,it} = \lambda_{2,i} f_{2,t} + \lambda_{2,i} + f_{2,t} + \eta_{2,it} + \mu_{2,i} \tag{31}$$

DGP9 differs from DGP7 in that the regressor $x_{l,it}$ for $l = 1, 2$ is only correlated with one factor $f_{j,t}$, the loadings $\lambda_{j,i}$ and the product $\lambda_{j,i} f_{j,t}$ for $j = 1, 2$, but is independent of the other factor and loadings. The unobserved factors follow a uniform distribution,

$$f_{j,t} \overset{iid}{\sim} \mathcal{U}(0, 2) \text{ for } j = 1, 2,$$

and the loadings follow a normal distribution,

$$\lambda_{j,i} \overset{iid}{\sim} \mathcal{N}(0, 4) \text{ for } j = 1, 2.$$

The regression errors are generated by an AR(1) process,

$$u_{it} = \rho u_{i,t-1} + \nu_{it} \text{ for } t = -49, ..., 0, 1, ...T,$$
$$\text{where } \rho = 0.5,\ \nu_{it} \sim \mathcal{N}(0, 1) \text{ and } u_{i,-50} = 0.$$

The first 50 time observations of u_{it} are discarded. The idiosyncractic components of the regressors are

$$\eta_{l,it} \overset{iid}{\sim} \mathcal{N}(0,1) \text{ and } \mu_{l,i} \overset{iid}{\sim} \mathcal{N}(0,1) \text{ for } l = 1, 2.$$