WORKING PAPER

A METHOD FOR EVALUATING THE RANK CONDITION FOR CCE ESTIMATORS

Ignace De Vos
Gerdie Everaert
Vasilis Sarafidis

April 2021
2021/1013

GHENT UNIVERSITY

Department of Economics

Faculty of Economics and Business Administration
Campus Tweekerken, St.-Pietersplein 5, 9000 Ghent - BELGIUM

A method for evaluating the rank condition for CCE estimators*

Ignace De Vos[1], Gerdie Everaert**[2], and Vasilis Sarafidis[3]

[1] *Lund University, Department of Economics*
[2] *Ghent University, Department of Economics*
[3] *BI Norwegian Business School, Department of Economics*

Abstract

This paper proposes a binary classifier to evaluate the so-called rank condition (RC), which is required for consistency of the Common Correlated Effects (CCE) estimator of Pesaran (2006). The RC postulates that the number of unobserved factors, m, is not larger than the rank of the unobserved matrix of average factor loadings, ϱ. When this condition fails, the CCE estimator is generally inconsistent. Despite the obvious importance of the RC, to date this condition could not be verified. The difficulty lies in that since the factor loadings are unobserved, ϱ cannot be evaluated or estimated directly. The key insight in the present paper is that ϱ can be established from the rank of the matrix of cross-sectional averages of *observables*. As a result, ϱ can be estimated consistently using procedures already available for determining the true rank of an unknown matrix. Similarly, m can be estimated consistently from the data using existing methods. A binary classifier that evaluates the RC is constructed by comparing the estimates of m and ϱ. The classifier correctly determines whether the RC is satisfied or not, with probability 1 as $(N, T) \to \infty$.

JEL classification: C13, C33, C52.

Keywords: Panel data, common factors, common correlated effects approach, rank condition.

[*] The authors thank participants of the 2017 and 2018 International Panel Data Conference for helpful comments and discussions. This paper also benefited from presentations at the 2018 Asian and European Meetings of the Econometric Society. Ignace De Vos gratefully acknowledges financial support from the Ghent University BOF research fund. Ignace De Vos and Gerdie Everaert further acknowledge financial support from the National Bank of Belgium. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the Flemish Supercomputer Center, funded by Ghent University; the Hercules Foundation; and the Economy, Science, and Innovation Department of the Flemish Government.

[**] Corresponding author: gerdie.everaert@ugent.be

1 Introduction

In a seminal paper, Pesaran (2006) put forward the Common Correlated Effects (CCE) approach for consistent estimation of panel data models with a multifactor error structure. The approach involves augmenting the regression model with "simple" (unweighted) cross-sectional averages (CSA) of the observables. Asymptotically, as the cross-sectional dimension (N) tends to infinity, the procedure aims to control for the unobserved common factors. Given the computational simplicity of the approach, involving least-squares, the CCE estimator has been highly popular, both in terms of extending it to several additional theoretical settings[1], as well as in terms of applying it to a large range of empirical areas[2].

Notwithstanding its simplicity, CCE comes at a cost. In particular, the CSA of the observables are valid proxies for the unobserved factors only if the number of factors, m, does not exceed the rank of the matrix of averaged factor loadings, ϱ. This restriction, known as the "rank condition" (RC), translates into the requirement that there must be at least as many observables holding linearly independent information about the unobserved factors, as the value of m. Westerlund and Urbain (2013) demonstrate that if the RC fails, then the CCE estimator is not consistent when the factor loadings are correlated with the regressors. More recently, Karabiyik et al. (2019) have shown that even when the factor loadings are uncorrelated with the regressors, failure of the RC leads to a lower rate of consistency for the CCE estimator.[3]

Despite the importance of the RC for the asymptotic properties of the CCE estimator, practitioners typically take it for granted. The main reason is that the matrix of average factor loadings is unobserved and therefore its rank cannot be evaluated or estimated directly.

This paper puts forward a binary classifier that evaluates the rank condition. The key insight is that the rank of the unobserved matrix of average factor loadings, ϱ, can be established from the rank of the matrix of CSA of the observables. As we shall show, this implies that ϱ can be estimated consistently using existing procedures developed for determining the true rank of an unknown matrix; see e.g. Camba-Mendez and Kapetanios (2009) and Al-Sadoon (2017) for an overview of this literature. Similarly, the number of factors, m, can be estimated from the data in a straightforward manner based on existing methods, such as those developed by Onatski (2010), Ahn and Horenstein (2013) and Kapetanios (2010), among many others. Comparing consistent estimates of m and

[1] See e.g. Kapetanios et al. (2011), Harding and Lamarche (2011), Su and Jin (2012), Chudik and Pesaran (2015), Everaert and De Groote (2016), Harding et al. (2018), Norkute et al. (2020) and De Vos and Everaert (2021), to mention a few.

[2] A recent search on Google Scholar indicated that the number of empirical applications based on CCE estimation currently exceeds one thousand.

[3] Note that in this case the standard two-way fixed effects estimator also remains consistent; see Sarafidis and Wansbeek (2012).

ϱ, $\widehat{m}$ and $\widehat{\varrho}$ respectively, the rank condition is deemed to be satisfied when the classifier $\widehat{RC} \equiv 1 - \mathbb{1}\{\widehat{\varrho} < \widehat{m}\} = 1$, where $\mathbb{1}\{\cdot\}$ is an indicator function that returns 1 when the argument inside the curly brackets holds true and 0 otherwise. The classifier is shown to be consistent, i.e. it determines correctly whether the rank condition is satisfied or not, with probability 1 as $(N, T) \rightarrow \infty$.

When the RC is violated for the standard CCE approach, one needs to augment the model with additional CSA that contain extra information about the factors. For instance, Pesaran et al. (2007) and Chudik and Pesaran (2015) advocate adding cross-sectional averages of external variables. Karabiyik et al. (2019) propose using external variables as weights, in order to construct additional, weighted-CSA. The weights are selected based on an Information Criterion (IC). The present paper contributes to this literature as well, by presenting alternative deterministic weights that arise by splitting the individual units into different groups, thus computing cluster-specific CSA.

In practice, it is not always clear which set of additional CSA to choose and whether the selected augmentations are sufficient to restore the rank condition for the augmented-CCE estimator. To address these issues, we put forward a strategy that combines the classifier proposed in the present paper and the IC criterion of Karabiyik et al. (2019). In particular, we first evaluate the RC for the standard CCE estimator, which makes use of simple (unweighted) CSA. If the RC is satisfied, there is no need to seek additional CSA. If the RC is found to be violated, we augment the model with additional CSA (motivated from the aforementioned potential choices), which are selected using the IC of Karabiyik et al. (2019). Subsequently, we evaluate the RC again for the augmented 'CCE_A' estimator. If the RC is still violated, more potential expansion CSA need to be sought, and so on. This strategy enables consistent CCE estimation of panel data models with a multifactor error structure, even in cases where the rank condition fails for the original CCE estimator.

We illustrate the practical relevance of our RC classifier by studying the effect of the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 on bank profitability. In particular, we analyse bank profitability conditional on several potential drivers, including bank size, capital adequacy, asset quality and liquidity. We use a random sample of 450 bank holding companies (BHCs) and banks, and we employ the CCE approach of Pesaran (2006) in order to account for macro-risk factors and common shocks, which hit the entire population of BHCs albeit with different intensities. To examine the impact of the Dodd-Frank Act, we estimate the model separately over two subperiods, namely 2006:Q1-2010:Q4 and 2011:Q1-2019:Q4. The RC classifier reveals that the rank condition holds for the standard CCE estimator in the second subperiod, but not in the first one. By augmenting the standard set of CSA using external variables, our procedure is able to restore the rank condition. This is important in the present study because the estimated effect of size on bank profitability is significantly lower when using the corresponding augmented CCE_A estimator. That is, the inconsistent CCE estimator overestimates the impact of bank size on profitability in the first sub-period.

The remainder of this paper is structured as follows. Section 2 introduces the model and the assumptions employed, and reviews the role of the rank condition. Section 3 develops a consistent classifier for evaluating the rank condition that underlies the CCE approach. Section 5 studies the finite sample properties of the proposed procedure. Section 6 illustrates our approach by examining the impact of the "Dodd-Frank Act" on bank profitability in the U.S. banking sector. A final section concludes. Proofs of theoretical results and additional simulation results are reported in Appendix A and Appendix B respectively.

In what follows we will use $\mathbf{A}^{\dagger}$ to denote the Moore-Penrose pseudo-inverse of the matrix $\mathbf{A}$, $rk(\mathbf{A})$ for its rank, $|\mathbf{A}|$ for the determinant and $\|\mathbf{A}\| = [tr(\mathbf{A}\mathbf{A}')]^{1/2}$ for its Euclidean (Frobenius) matrix norm. A vec(.) denotes the vectorization operation. Finally, $\lfloor a \rfloor$ ($\lceil a \rceil$) is the floor (ceiling) function, which yields the largest (smallest) integer less than (greater than) or equal to a.

2 A multi-factor panel data model and the CCE approach

2.1 Model and assumptions

We study the following linear regression model with unobserved common factors

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{F}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\mathbf{y}_i = [y_{i1}, \ldots, y_{iT}]'$ denotes a $T \times 1$ vector of observations on the dependent variable for individual i, $\mathbf{X}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}]'$ denotes a $T \times K$ matrix of covariates, where $\mathbf{x}_{it}$ is $K \times 1$, and $\boldsymbol{\beta}$ is a $K \times 1$ vector of unknown parameters of interest with $\|\boldsymbol{\beta}\| < \infty$. The error term is composite, such that $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_T]'$ denotes a $T \times m$ matrix of unobserved common factors, where $\mathbf{f}_t$ is $m \times 1$, and $\boldsymbol{\lambda}_i$ denotes an $m \times 1$ vector of factor loadings. Finally, $\boldsymbol{\varepsilon}_i = [\varepsilon_{i1}, \ldots, \varepsilon_{iT}]'$ is a $T \times 1$ vector of purely idiosyncratic disturbances.

Following Pesaran (2006), we assume that the covariates are also subject to a common factor structure, such that the data generating process (DGP) for $\mathbf{X}_i$ is given by

$$\mathbf{X}_i = \mathbf{F}\boldsymbol{\Gamma}_i + \mathbf{V}_i, \tag{2}$$

where $\boldsymbol{\Gamma}_i$ denotes an $m \times K$ matrix of factor loadings, and $\mathbf{V}_i = [\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT}]'$ is a $T \times K$ matrix of idiosyncratic errors.

Replacing $\mathbf{X}_i$ in Eq. (1) by the expression in Eq. (2), and stacking the observables into a $T \times (K+1)$ matrix $\mathbf{Z}_i = [\mathbf{y}_i, \mathbf{X}_i] \equiv [\mathbf{z}_{i1}, ..., \mathbf{z}_{iT}]'$, yields

$$\mathbf{Z}_i = \mathbf{F}\mathbf{C}_i + \mathbf{U}_i, \tag{3}$$

where $\mathbf{C}_i = [\boldsymbol{\delta}_i, \boldsymbol{\Gamma}_i]$ is of order $m \times (K+1)$ with $\boldsymbol{\delta}_i = \boldsymbol{\lambda}_i + \boldsymbol{\Gamma}_i\boldsymbol{\beta}$, and $\mathbf{U}_i = [\boldsymbol{\varepsilon}_i + \mathbf{V}_i\boldsymbol{\beta}\,, \mathbf{V}_i]$. In what follows, it is important to note that $\mathbf{C}_i$ can be written as $\mathbf{C}_i = \widetilde{\mathbf{C}}_i\mathbf{B}$, with

$$\widetilde{\mathbf{C}}_i = [\boldsymbol{\lambda}_i, \boldsymbol{\Gamma}_i]\,, \qquad \mathbf{B} = \begin{bmatrix} 1 & \mathbf{0}_{1\times K} \\ \boldsymbol{\beta} & \mathbf{I}_K \end{bmatrix}. \tag{4}$$

Therefore, since $\mathbf{B}$ is always a full rank square matrix, the rank of $\mathbf{C}_i$ is solely determined by the matrix of factor loadings $\widetilde{\mathbf{C}}_i$.

The following assumptions are made throughout the paper:
Assumption 1. *(Idiosyncratic errors) ε_{it} and $\mathbf{v}_{it}$ are mean zero, covariance-stationary variables that are i.i.d. across i, with $E(\varepsilon_{it}^4) < \infty$ and $E(\|\mathbf{v}_{it}\|^4) < \infty$ for all i and t.*
Assumption 2. *(Common factors) $\mathbf{f}_t$ is covariance-stationary with $E(\|\mathbf{f}_t\|^4) < \infty$ and absolute summable autocovariances, such that $T^{-1}\mathbf{F}'\mathbf{F} \to \boldsymbol{\Sigma}_F$ as $T \to \infty$, where $\boldsymbol{\Sigma}_F$ is positive definite.*
Assumption 3. *(Factor loadings) $\widetilde{\mathbf{C}}_i$ is generated according to*

$$\widetilde{\mathbf{C}}_i = \widetilde{\mathbf{C}} + \boldsymbol{\Xi}_i; \qquad \boldsymbol{\xi}_i \sim i.i.d.(\mathbf{0}_{m(K+1)}, \boldsymbol{\Omega}_\xi), \tag{5}$$

where $\widetilde{\mathbf{C}} = E(\widetilde{\mathbf{C}}_i) \equiv [\boldsymbol{\lambda}, \boldsymbol{\Gamma}]$ such that $\left\|\widetilde{\mathbf{C}}\right\| < \infty$, $\boldsymbol{\xi}_i = vec(\boldsymbol{\Xi}_i)$, $\boldsymbol{\Omega}_\xi = E(\boldsymbol{\xi}_i\boldsymbol{\xi}_i')$ with $\|\boldsymbol{\Omega}_\xi\| < \infty$. In addition, $\frac{1}{N}\sum_{i=1}^N \mathbf{C}_i\mathbf{C}_i' \to \boldsymbol{\Sigma}_C$ as $N \to \infty$, with $\boldsymbol{\Sigma}_C$ positive definite.
Assumption 4. *(Dataset dimension) $T \geq m$, with m fixed and finite.*
Assumption 5. *(Independence) $\mathbf{f}_t, \varepsilon_{is}, \mathbf{v}_{jl}, \boldsymbol{\xi}_h$ are mutually independent for all t, i, s, j, l, h.*

The setup described by the DGP in Eq. (3) together with Assumptions 1-5, is similar to that in Pesaran (2006) but deviates in the following respects. First, we focus on a model with homogeneous slope coefficients and without fixed effects. This is for ease of exposition only, as the results below also follow through under the assumption of independent random coefficients with a common mean, as made in Pesaran (2006). Second, following Westerlund and Urbain (2013) and Karabiyik et al. (2019), Assumption 3 generalizes Pesaran (2006) by allowing $\boldsymbol{\lambda}_i$ and $\boldsymbol{\Gamma}_i$ to be correlated across i. Third, we introduce more explicit regularity conditions on the factors and their loadings compared to what is typically the case in the CCE literature. In particular, the non-central second moments are assumed to converge to a positive definite matrix. Such regularity conditions are common in the factor literature (see e.g. Bai and Ng, 2002) and allow us to consistently estimate m.

2.2 CCE and the rank condition

Given that $\mathbf{F}$ enters into the data generating process of both $\mathbf{y}_i$ and $\mathbf{X}_i$, and since $\boldsymbol{\lambda}_i$ and $\boldsymbol{\Gamma}_i$ are allowed to be mutually correlated, failure to account for the common factor component leads to endogeneity of $\mathbf{X}_i$. Therefore, standard panel data estimators, such as the two-way fixed and random effects estimators, fail to be consistent for the parameters of interest, $\boldsymbol{\beta}$. The key idea of the CCE approach is to replace the unobserved factors with CSA of the observables in Eq. (3).

In particular, taking sample averages over i in Eq. (3), we obtain

$$\underset{T\times(K+1)}{\overline{\mathbf{Z}}} = \underset{T\times m}{\mathbf{F}}\ \underset{m\times(K+1)}{\overline{\mathbf{C}}} + \underset{T\times(K+1)}{\overline{\mathbf{U}}}, \tag{6}$$

where barred variables denote CSA as in $\overline{\mathbf{Z}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i$.

Under Assumptions 1-5 it is straightforward to show that $\overline{\mathbf{U}} = O_p(N^{-1/2})$ and $\overline{\mathbf{C}} = \mathbf{C} + O_p(N^{-1/2})$, where $\mathbf{C} = E(\mathbf{C}_i)$. As a result, $\overline{\mathbf{Z}}$ converges to a linear combination of the m unobserved common factors, i.e.,

$$\overline{\mathbf{Z}} = \mathbf{FC} + \mathbf{F}(\overline{\mathbf{C}} - \mathbf{C}) + \overline{\mathbf{U}} = \mathbf{FC} + O_p(N^{-1/2}). \tag{7}$$

Suppose that $\mathbf{C}$ has full rank, such that $\mathbf{CC}'$ is invertible and bounded by Assumption 3, then post-multiplying Eq. (7) by $\mathbf{C}'$ and solving for $\mathbf{F}$ yields

$$\mathbf{F} = \overline{\mathbf{Z}}\mathbf{C}'(\mathbf{CC}')^{-1} + O_p(N^{-1/2}). \tag{8}$$

Thus, in finite samples the common factor component can be controlled for in estimation by re-writing the original model in Eq. (1) as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{F}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i = \mathbf{X}_i\boldsymbol{\beta} + \overline{\mathbf{Z}}\boldsymbol{\lambda}_i^* + \boldsymbol{\varepsilon}_i^*, \tag{9}$$

where $\boldsymbol{\lambda}_i^* = \overline{\mathbf{C}}^+\boldsymbol{\lambda}_i$, $\overline{\mathbf{C}}^+ = \overline{\mathbf{C}}'(\overline{\mathbf{C}}\overline{\mathbf{C}}')^{-1}$ and $\varepsilon_i^* = \boldsymbol{\varepsilon}_i - \overline{\mathbf{U}}\overline{\mathbf{C}}^+\boldsymbol{\lambda}_i$.

The corresponding pooled CCE (CCEP) estimator for $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{M}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{M}\mathbf{y}_i, \tag{10}$$

where $\mathbf{M} = \mathbf{I}_T - \overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\overline{\mathbf{Z}})^{+}\overline{\mathbf{Z}}'$.[4]

The CCE approach crucially relies upon the assumption that $\mathbf{C}$ is full rank. This restriction, known as the "rank condition" (RC), can be expressed as

$$\varrho = m \tag{11}$$

where we defined $\varrho = rk(\mathbf{C})$. As pointed out earlier, when the RC fails, such that $\varrho < m$, the CCEP estimator is in general not consistent as the unobserved factor space will not be controlled for. Unfortunately, there exist several cases where the RC may fail in practice. For instance, when $m > K+1$, i.e. the number of factors is larger than the number

[4] When the model contains fixed effects, then one may set $\mathbf{M} = \mathbf{I}_T - \overline{\mathbf{H}}(\overline{\mathbf{H}}'\overline{\mathbf{H}})^{+}\overline{\mathbf{H}}'$, where $\overline{\mathbf{H}} = [\boldsymbol{\iota}_T, \overline{\mathbf{Z}}]$ and $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones.

of CSA employed, it is straightforward to see that $\varrho \leq min\{m, K+1\} = K+1 < m$. Intuitively, re-writing Eq. (7) in vector form at time t, we have

$$\bar{\mathbf{z}}_t = \begin{bmatrix} \bar{y}_t \\ \bar{\mathbf{x}}_t \end{bmatrix} = \mathbf{C}'\mathbf{f}_t + O_p(N^{-1/2}).$$

It is impossible to solve the above system of $K+1$ equations in terms of $m > K+1$ unknown factors $\mathbf{f}_t$.

Note that $K+1 \geq m$ is a necessary but not a sufficient condition for the RC to hold true. For example, certain columns of $\overline{\mathbf{Z}}$ can be asymptotically uninformative because the corresponding observables: (i) do not load on the common factors (e.g. $\mathbf{\Gamma}_i = \mathbf{0}$); (ii) have factor loadings that average out (e.g. $\overline{\mathbf{\Gamma}} = O_p(N^{-1/2})$); and (iii) do not hold information on the common factors that is distinct from the information already provided by other observables. In all these cases, the number of informative observables in $\overline{\mathbf{Z}}$, measured by ϱ, can be lower than m.

3 Evaluation of the rank condition

Despite the importance of the RC for the asymptotic properties of the CCE estimator, it is typically taken for granted in practice. The main reason arguably is that the matrix of CSA of the factor loadings, $\mathbf{C}$, is unobserved. Therefore the rank of this matrix cannot be evaluated or estimated directly.

The key insight of this paper is that the rank of $\mathbf{C}$ can be determined from the observed matrix of CSA, $\overline{\mathbf{Z}}$. To see this, recall from Eq. (7) that $\overline{\mathbf{Z}} = \mathbf{FC} + O_p(N^{-1/2})$. It follows that so long as $T \geq m$ (Assumption 4),[5] then under the maintained assumptions the rank of $\overline{\mathbf{Z}}$ asymptotically equals that of $\mathbf{C}$. This result is summarized in the following proposition:

Proposition 1. *Let* $\overline{\mathbf{Z}} = \mathbf{FC} + O_p(N^{-1/2})$. *As* $N \to \infty$,

$$rk(\overline{\mathbf{Z}}) \overset{a.s.}{=} rk\left(\mathbf{FC}\right) = rk\left(\mathbf{C}'\mathbf{F}'\right) = rk\left(\mathbf{C}'\right) = rk\left(\mathbf{C}\right) = \varrho. \tag{12}$$

The third equality follows from the fact that $\mathbf{F}'$ has full row rank by Assumption 2 and $T > m$.[6]

Proposition 1 implies that estimating $rk(\overline{\mathbf{Z}})$ is asymptotically equivalent to estimating ϱ. Hence, one can use the observed $\overline{\mathbf{Z}}$ to infer the rank of $\mathbf{C}$, without observing $\mathbf{C}$ itself.

Once a consistent estimate of ϱ is obtained, the RC can be evaluated by direct comparison of this value with a consistent estimate for m. The latter can be determined from the

[5] This condition trivially satisfied when we consider large T asymptotics, since m is fixed.

[6] See pg. 85 in Abadir and Magnus (2005) for more details.

observed data in a straightforward manner based on existing literature; see e.g. Bai and Ng (2002), Alessi et al. (2010), Onatski (2010) and Ahn and Horenstein (2013), among many others.

The following two sections provide details for consistent estimation of ϱ and m. Section 3.3 puts forward a binary classifier that evaluates the rank condition correctly with probability 1 as $(N, T) \to \infty$. Section 4 discusses a strategy for obtaining a consistent CCEP estimator when the RC fails.

3.1 Consistent estimation of ϱ

Building on the result of Proposition 1, the rank of $\mathbf{C}$ can be determined based on the rank of $\overline{\mathbf{Z}}$. The problem of testing for the rank of a general matrix is long standing in the econometrics literature, and many methods have been suggested; see e.g. Cragg and Donald (1997), Robin and Smith (2000) and Camba-Mendez and Kapetanios (2009).[7] A closely-related problem, which is more relevant in the present context, is the problem of consistently estimating the true rank of a matrix. The two dominant approaches considered thus far are based on either sequential testing procedures, or information criteria. Camba-Mendez and Kapetanios (2009) provide an overview of these approaches and conclude that sequential testing procedures have an advantage over standard information criteria methods under several modeling scenarios. Therefore, in the remainder of this section, we closely follow the sequential testing procedure advocated by Robin and Smith (2000). This procedure is straightforward to implement and relies on relatively mild assumptions, in that it does not require the variance-covariance of the estimator of the unknown matrix $\mathbf{FC}$ to be full rank, or its rank to be known.

A complicating factor in the present paper relative to the method of Robin and Smith (2000) is that therein the dimensions of the matrix for which the rank is to be estimated are fixed as the sample size grows. In contrast, here the matrix of interest $\overline{\mathbf{Z}}$ is of order $T \times n$ such that the number of rows (and so, the number of eigenvalues of $\overline{\mathbf{Z}}\overline{\mathbf{Z}}'$) increases with T. To circumvent this issue, we introduce a narrow matrix $\mathbf{\Psi}$ of order $n \times T$, such that $rk(\mathbf{\Psi}\overline{\mathbf{Z}}) = rk(\overline{\mathbf{Z}})$. That is, $\mathbf{\Psi}$ has the role of reducing the dimensionality of $\overline{\mathbf{Z}}$ without altering its rank. The following assumption is employed:

Assumption 6. *(Dimensionality reduction matrix)* $\mathbf{\Psi}$ *satisfies*

$$(i) \quad \|\mathbf{\Psi F}\| = O_p(1); \qquad (ii) \quad \left\|\mathbf{\Psi}\overline{\mathbf{U}}\right\| = O_p(N^{-1/2});$$
$$(iii)\ \sqrt{N}\text{vec}(\mathbf{\Psi}\overline{\mathbf{Z}} - \mathbf{\Psi FC}) \to^{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, \mathbf{\Omega}\right).$$

Assumption 6(i) implies that the entries of $\mathbf{\Psi}$ are sufficiently bounded. Assumption 6(ii) states that $\mathbf{\Psi}$ is asymptotically uncorrelated with $\overline{\mathbf{U}}$, the error term in Eq. (6). Assumption 6(iii) ensures that, by application of a suitable central limit theorem, $\mathbf{\Psi}\overline{\mathbf{Z}}$ remains

[7] See Al-Sadoon (2017) for an in-depth study of the relation among various tests of rank.

$\sqrt{N}$-consistent for $\mathbf{\Psi FC}$ and asymptotically normally distributed as $N \to \infty$. This assumption is identical to Assumption 2.2. in Robin and Smith (2000), except that it is imposed on $\mathbf{\Psi}\overline{\mathbf{Z}}$ rather than $\overline{\mathbf{Z}}$ itself.

Given the above, we propose estimating the rank of $\mathbf{\Psi}\overline{\mathbf{Z}}$, an $n \times n$ matrix, by sequentially testing the null hypothesis $H_0 : \varrho = \varrho^*$ against the alternative $H_a : \varrho > \varrho^*$, using the following statistic:

$$\tau = N \sum_{\ell=\varrho^*+1}^{n} \lambda_\ell(\mathbf{A}), \tag{13}$$

where $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ are the ordered eigenvalues of the $n \times n$ matrix $\mathbf{A} \equiv \mathbf{\Psi}\overline{\mathbf{Z}}\overline{\mathbf{Z}}'\mathbf{\Psi}'$. The procedure is implemented sequentially for $\varrho^* = 0, \ldots, n-1$ and the estimated rank $\widehat{\varrho}$ corresponds to the smallest value of ϱ^* for which the null hypothesis is not rejected. Under the null, τ has a limiting distribution which is a weighted sum of independent χ^2_1 variables, with weights given by the $(n-\varrho^*)^2$ largest eigenvalues of $(\mathbf{D}'_{\varrho^*} \otimes \mathbf{R}'_{\varrho^*})\mathbf{\Omega}(\mathbf{D}_{\varrho^*} \otimes \mathbf{R}_{\varrho^*})$, where $\mathbf{D}_{\varrho^*}$ and $\mathbf{R}_{\varrho^*}$ denote the eigenvectors corresponding to the $n-\varrho^*$ smallest eigenvalues of $\overline{\mathbf{Z}}'\mathbf{\Psi}'\mathbf{\Psi}\overline{\mathbf{Z}}$ and $\mathbf{A}$, respectively. This is summarized in the following proposition:

Proposition 2. *Suppose that Assumption 6 holds true. Then, as $N \to \infty$,*

$$\tau \to^{\mathcal{L}} \sum_{\ell=1}^{(n-\varrho^*)^2} \varpi_\ell \mathcal{Z}^2_\ell, \tag{14}$$

where ϖ_ℓ is the ℓth largest eigenvalue of $(\mathbf{D}'_{\varrho^*} \otimes \mathbf{R}'_{\varrho^*})\mathbf{\Omega}(\mathbf{D}_{\varrho^*} \otimes \mathbf{R}_{\varrho^*})$ and $\{\mathcal{Z}_\ell\}_{\ell=1}^{(n-\varrho^*)(n-\varrho^*)}$ are independent standard normal variates such that $\mathcal{Z}^2_\ell \sim \chi^2_1$ is independent across ℓ.

The proof of Proposition 2 follows from similar arguments as in Robin and Smith (2000), mutatis mutandis. We omit the proof to save space.

Note that although $\mathbf{\Omega}$ is unknown, it can be estimated consistently based on the following expression[8]:

$$\widehat{\mathbf{\Omega}} = \frac{1}{N}\sum_{i=1}^{N} \text{vec}(\mathbf{\Psi Z}_i - \mathbf{\Psi}\overline{\mathbf{Z}}_w)\text{vec}(\mathbf{\Psi Z}_i - \mathbf{\Psi}\overline{\mathbf{Z}})'. \tag{15}$$

Remark 3.1. In order to consistently estimate ϱ, the significance level α_N employed in the testing sequence needs to decrease as N grows. This is because α_N is the probability of over-estimating the true rank, $P(\widehat{\varrho} > \varrho)$, which must tend to zero for $\widehat{\varrho}$ to be consistent. Hence, α_N ought to decrease sufficiently fast with N to limit the number of times ϱ is over-estimated, but not too fast, as this would result in severe under-estimation of the true rank when N is small. More specifically, Robin and Smith (2000) show that $\alpha_N = o(1)$ and $-N^{-1}\ln\alpha_N = o(1)$ are sufficient for consistency. We suggest using

[8] $\mathbf{\Omega}$ can also be estimated using bootstrap techniques. When the model contains fixed constants, $\mathbf{Z}_i$ should be time-demeaned, i.e., pre-multiplied with $\mathbf{Q} = \mathbf{I}_T - \boldsymbol{\iota}_T\boldsymbol{\iota}'_T/T$.

$\alpha_N = \alpha c N^{-1/\gamma}$. This way, for a given choice of α and γ, the level of α_N in samples with small N can be controlled by setting $c > 1$, whereas the speed at which α_N decreases in N can be governed by setting $\gamma > 0$. For instance, choosing $\alpha = 5\%$ and setting $c = 20$ and $\gamma = 1$ fixes the nominal significance level to 5% for $N = 20$ and lets it decrease at rate N. Given that over-estimating ϱ may lead to falsely concluding that the rank condition is satisfied, we prefer a high rate of decrease with N in order to be conservative (i.e., requiring strong evidence against the null before rejecting it in favor of a higher rank estimate) and cap the probability of erroneously concluding that the rank condition holds true.

In practice, there exist several potential choices for $\boldsymbol{\Psi}$. One possible (stochastic) choice is to draw $\boldsymbol{\Psi}$ from the standard normal distribution. The following theorem confirms that such choice is rank-preserving. Moreover, pre-multiplication of $\overline{\mathbf{Z}}$ by $T^{-1/2}\boldsymbol{\Psi}$ ensures that the product also adheres to the required conditions above.

Theorem 1. *Let $T > n$ and $\boldsymbol{\Psi}$ be a $n \times T$ random matrix with i.i.d. standard normal entries.*

(i) It holds that

$$rk(\boldsymbol{\Psi}\overline{\mathbf{Z}}) = rk(\overline{\mathbf{Z}}),$$

that is, the rank of $\overline{\mathbf{Z}}$ is preserved by the random projection $\boldsymbol{\Psi}$.

(ii) Under Assumptions 1-3, 5 it follows that

$$T^{-1/2}\boldsymbol{\Psi}\overline{\mathbf{Z}} = T^{-1/2}\boldsymbol{\Psi}\mathbf{F}\mathbf{C} + O_p(N^{-1/2}),$$

where $\left\|T^{-1/2}\boldsymbol{\Psi}\mathbf{F}\mathbf{C}\right\| = O_p(1)$.

The proof of Theorem 1 is in Appendix A.

Remark 3.2. An alternative stochastic choice would be to set $\boldsymbol{\Psi} = \overline{\mathbf{Z}}'/T$. However, this choice is ruled out because even though $\boldsymbol{\Psi}\overline{\mathbf{Z}} = \overline{\mathbf{Z}}'\overline{\mathbf{Z}}/T$ is stochastically bounded and has the same rank as $\overline{\mathbf{Z}}$, it does not have an asymptotic normal distribution, i.e. it violates Assumption 6(*iii*).

Alternatively, the choice for $\boldsymbol{\Psi}$ can be deterministic. In particular, since n time periods contain the same amount of information on the rank of $\mathbf{C}$ as do T observations, an obvious candidate is to let $\boldsymbol{\Psi} = [\mathbf{0}_{n\times(T-n)}, \mathbf{I}_n]$, which considers only the last n time periods in $\overline{\mathbf{Z}}$. Moreover, one can also fold over (i.e. sum every n rows) and average the $\overline{\mathbf{Z}}$ matrix over time. This corresponds to setting $\boldsymbol{\Psi} = \frac{1}{\lceil T/n\rceil}[\boldsymbol{\iota}'_{\lceil T/n\rceil} \otimes \mathbf{I}_n]\mathbf{I}_{\lceil T/n\rceil n,T}$, where $\boldsymbol{\iota}_{\lceil T/n\rceil}$ is a $\lceil T/n\rceil \times 1$ vector of ones. The rank-preserving properties of these projections have also been verified in (unreported) simulations.

3.2 Consistent estimation of *m*

There is a substantial literature on estimating the number of factors from observed data. Briefly speaking, existing methods involve one of the following three approaches:

looking at differences or ratios of adjacent eigenvalues (e.g. Onatski (2010) and Ahn and Horenstein (2013)), specifying threshold functions to separate bounded from unbounded eigenvalues (e.g. Bai and Ng (2002) and Alessi et al. (2010)), or using sequential tests that determine which eigenvalues are unbounded (Kapetanios, 2010; Trapani, 2018). Preliminary simulation evidence conducted for this paper shows that the Growth Ratio (GR) by Ahn and Horenstein (2013) performs well in finite samples and outperforms other estimators.[9] Therefore, in what follows we propose estimating m using the GR statistic.

In particular, let $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_N]$ denote a $T \times (K+1)N$ matrix, where $\mathbf{Z}_i$ (defined in Eq. (3)) collects all observables for individual i in a $T \times (K+1)$ matrix. Also, let m_{max} denote the maximum value of m considered in estimation, such that $m_{max} \geq m$. We define

$$\widehat{m} = \underset{j \in \{1,\ldots,m_{max}\}}{\arg\max}\ GR(j); \quad GR(j) = \frac{ln(V(j-1)/V(j))}{ln(V(j)/V(j+1))}, \tag{16}$$

where $V(j) = \sum_{k=j+1}^{h} \lambda_j \left(\mathbf{Z}\mathbf{Z}'/NT\right)$ with $h = min\{T, N(K+1)\}$, and $\lambda_j \left(\mathbf{Z}\mathbf{Z}'/NT\right)$ denotes the jth largest eigenvalue of $(\mathbf{Z}\mathbf{Z}'/NT)$.

The above GR statistic is easy to compute because it involves maximizing the "growth ratio" of two adjacent eigenvalues arranged in descending order. The main intuition is that the growth ratios of two adjacent eigenvalues of $\mathbf{Z}\mathbf{Z}'/NT$ are asymptotically bounded, except for the growth ratio involving the mth and $(m+1)$th eigenvalues, which diverges to infinity.

Under regularity conditions implied by Assumptions 1-5, Ahn and Horenstein (2013) show that

$$lim_{min\{N,T\}\to\infty} Pr\left(\widehat{m} = m\right) = 1, \tag{17}$$

for any $m_{max} \in \{m, (d^c min\{N, T\}) - m - 1\}$, where $d^c \in (0, 1]$.

Remark 3.3. In exactly the same way as described above, the number of factors can also be estimated based on the $T \times T$ matrix $\mathbf{Y}\mathbf{Y}'/NT$, where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ is of dimension $T \times N$. However, since both $\mathbf{y}_i$ and $\mathbf{X}_i$ share the same factors by assumption, it is natural to combine them together in order to increase the information set used to construct proxies for $\mathbf{F}$. This strategy is in line with the rationale behind the CCE approach, which involves solving a system of equations, such that Eq. (3) includes LHS variables (observables) that are solely driven by a common factor component and purely idiosyncratic noise. Moreover, this strategy is consistent with Westerlund and Urbain (2015), who also estimate factors based on $\mathbf{Z}\mathbf{Z}'/NT$.

[9] Juodis and Sarafidis (2018) provide additional evidence that confirms the good performance of the GR statistic in finite samples.

3.3 A consistent classifier for the rank condition

Given consistent estimates of the rank of $\mathbf{C}$ and the number of factors, $\widehat{\varrho}$ and $\widehat{m}$ respectively, the rank condition is deemed to be violated when $\widehat{\varrho} < \widehat{m}$. In particular, we define the following classifier:

$$\widehat{RC} \equiv 1 - \mathbb{1}\{\widehat{\varrho} < \widehat{m}\}, \tag{18}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function that returns 1 when the argument inside the curly brackets holds true, and 0 otherwise. Hence, if $\widehat{RC} = 1$ the rank condition is considered to be satisfied, whereas $\widehat{RC} = 0$ indicates that (11) may be violated. The definition in (18) shows that we also take $\widehat{\varrho} > \widehat{m}$ as a sign that (11) is satisfied.[10]

The following proposition summarizes the asymptotic properties of the proposed classifier:

Proposition 3. *Let Assumptions 1-6 hold true. Suppose also that ϱ is determined based on the sequential testing procedure outlined in Section 3.1, with $\alpha_N = o(1)$, and $-N^{-1}\ln\alpha_N = o(1)$, and m is determined by Eq. (16). Then, as $(N, T) \to \infty$,*

$$Pr\left[\left(\widehat{RC} = 1|\varrho = m\right) \cup \left(\widehat{RC} = 0|\varrho < m\right)\right] \to 1. \tag{19}$$

That is, the probability that the classifier correctly identifies whether the rank condition is satisfied or not, converges to unity. The result follows directly from the consistency of $\widehat{\varrho}$ as $N \to \infty$ under Assumptions 1-6, given an appropriate rate of decay for α_N, and the consistency of $\widehat{m}$ as $(N, T) \to \infty$.

4 What if the rank condition is violated?

When $\widehat{RC} = 0$, the standard CCE estimator is inconsistent in general, unless the regressors are uncorrelated with the unobserved factor loadings. To circumvent this problem, one may seek to restore the RC by means of augmenting the existing model with additional CSA. There are several potential strategies available for this purpose.

For example, Pesaran et al. (2007) and Chudik and Pesaran (2015) advocate expanding $\overline{\mathbf{Z}}$ by adding cross-sectional averages of external variables. This practice requires that these variables load on the same set of factors $\mathbf{F}$ that operate in $\mathbf{Z}_i$, but otherwise have no relation to the dependent variable. To illustrate, consider a setting where $m > K + 1$ so that the rank condition is violated for $\overline{\mathbf{Z}}$. Let $\mathbf{Z}_i^{(e)}$ be the $T \times K_e$ matrix gathering the exogenous covariates, given by

$$\mathbf{Z}_i^{(e)} = \mathbf{F}\mathbf{C}_i^{(e)} + \boldsymbol{\epsilon}_i^{(e)}, \tag{20}$$

[10] $\widehat{\varrho} > \widehat{m}$ can only occur in finite samples due to estimation error but not at the population level.

where $\mathbf{C}_i^{(e)}$ denotes an $m \times K_e$ matrix of factor loadings with finite mean $\mathbf{C}^{(e)}$, and $\boldsymbol{\epsilon}_i^{(e)}$ is the $T \times K_e$ matrix of errors. Assuming that the components of this DGP also satisfy Assumptions 1-3 and 5, the augmented matrix of CSA, $\overline{\mathbf{Z}}_A = \left[\overline{\mathbf{Z}}, \overline{\mathbf{Z}}^{(e)}\right]$, may satisfy the rank condition, because it can be written as

$$\overline{\mathbf{Z}}_A = \mathbf{F}\left[\overline{\mathbf{C}}, \overline{\mathbf{C}}^{(e)}\right] + \left[\overline{\mathbf{U}}, \overline{\boldsymbol{\epsilon}}^{(e)}\right] = \mathbf{F}\mathbf{C}_A + O_p(N^{-1/2}), \tag{21}$$

where $\mathbf{C}_A = [\mathbf{C}, \mathbf{C}^{(e)}]$. Given that $\mathbf{Z}_i^{(e)}$ loads on the same set of factors $\mathbf{F}$, the augmented loading matrix $\mathbf{C}_A$ is now of order $m \times (1 + K + K_e)$. Therefore, this can restore the RC provided that $m \leq 1 + K + K^{(e)}$ and $\mathbf{C}^{(e)}$ is also sufficiently distinct from $\mathbf{C}$.

An alternative idea is to make use of external variables as additional weights, in order to construct *weighted* CSA. Such an approach has been recently advocated by Juodis and Sarafidis (2020), Fan and Liao (2020), Juodis and Sarafidis (2021) and, in the present context of CCE estimation, by Karabiyik et al. (2019).

To illustrate, let w_i denote an external, time-invariant variable.[11] Multiplying Eq. (3) by w_i and summing over i yields

$$\underset{T\times(K+1)}{\overline{\mathbf{Z}}_w} = \underset{T\times m}{\mathbf{F}} \underset{m\times(K+1)}{\overline{\mathbf{C}}_w} + \underset{T\times(K+1)}{\overline{\mathbf{U}}_w}, \tag{22}$$

where $\overline{\mathbf{Z}}_w = \sum_{i=1}^N \mathbf{Z}_i w_i$, $\overline{\mathbf{C}}_w = \sum_{i=1}^N \mathbf{C}_i w_i$, and $\overline{\mathbf{U}}_w = \sum_{i=1}^N \mathbf{U}_i w_i$. As shown by Karabiyik et al. (2019), when $\mathbf{C}_i$ and w_i are correlated, but $\mathbf{U}_i$ and w_i are not, then $\overline{\mathbf{Z}}_w = \mathbf{F}\overline{\mathbf{C}}_w + O_p(N^{-1/2})$ and $\overline{\mathbf{C}}_w$ converges to a nonzero matrix.[12] If $\overline{\mathbf{C}}_w$ is also sufficiently distinct from $\overline{\mathbf{C}}$, the obtained $\overline{\mathbf{Z}}_w$ provides new (i.e. rank increasing) information on $\mathbf{F}$, and the rank of the augmented matrix $\overline{\mathbf{Z}}_A = [\overline{\mathbf{Z}}, \overline{\mathbf{Z}}_w]$ is increased. As the authors point out, w_i effectively acts as an instrument for $\mathbf{C}_i$, and multiple w_i can be combined in an attempt to restore the RC.[13]

Lastly, one can also employ deterministic averaging weights, such as binary indicators that give rise to group-specific cross-sectional averages. For example, in a panel of countries, individual units may be classified as developed, emerging and developing economies; in a panel of firms, units may be grouped according to their size or sector; and so on. In many cases, such group memberships are known and the group-specific averages can be more informative factor proxies than the simple (overall) average.

The rich variety of potential expansions suggested above brings about two important issues. First of all, the issue of how to choose CSA from a set of candidate expansions. This is particularly relevant when some weights violate Assumption 2 in Karabiyik et al.

[11] For example, Karabiyik et al. (2019) estimate a gravity equation of bilateral trade flows and construct weights based on different measures of trade cost.

[12] This property is also utilised by Juodis and Sarafidis (2021), who propose the use of aggregation weights in the context of GMM estimation in panels with T fixed or large.

[13] See Section 2 in Karabiyik et al. (2019) for the formal set of assumptions required to ensure the validity of such weights.

(2019), or some candidate CSA load on different factors than those in the regression model. Second, given a set of selected additional CSA, one still needs to check whether the rank condition is satisfied for the augmented-CCE estimator.

In order to tackle the first issue, Karabiyik et al. (2019) have proposed an information criterion (IC) selection procedure. To illustrate, let $\overline{\mathbf{Z}}_+$ be the matrix of available expansions

$$\overline{\mathbf{Z}}_+ = \{\overline{\mathbf{Z}}_+^{(1)}, \overline{\mathbf{Z}}_+^{(2)}, \overline{\mathbf{Z}}_+^{(3)}\} \tag{23}$$

where (say) $\overline{\mathbf{Z}}_+^{(1)} = \overline{\mathbf{Z}}^{(e)}$ contains CSA of exogenous variables, $\overline{\mathbf{Z}}_+^{(2)} = \overline{\mathbf{Z}}_{w_1}$ contains a matrix of CSA arising from a specific weight w_1, and similarly $\overline{\mathbf{Z}}_+^{(3)} = \overline{\mathbf{Z}}_{w_2}$ for a weight w_2. The appropriate set of expansion CSA can be selected from $\overline{\mathbf{Z}}_+$ by minimizing

$$\boldsymbol{\ell}^* = \arg\min_{\boldsymbol{\ell}} IC(\boldsymbol{\ell}) \tag{24}$$

where

$$IC(\boldsymbol{\ell}) = \ln|\Sigma_{i=1}^N \mathbf{Z}_i' \mathbf{M}_A^{(\boldsymbol{\ell})} \mathbf{Z}_i / NT| + g(n); \qquad g(n) = n(K+1)\frac{\ln\left(\min\{N, \sqrt{T}\}\right)}{\min\{N, \sqrt{T}\}}, \tag{25}$$

and $\boldsymbol{\ell} = \{\ell_1, \ell_2, ...\}$ gathers the indices of the considered expansions from $\overline{\mathbf{Z}}_+$ such that for (say) $\boldsymbol{\ell} = \{\ell_1, \ell_3\}$, $\overline{\mathbf{Z}}_A^{(\boldsymbol{\ell})} = [\overline{\mathbf{Z}}, \overline{\mathbf{Z}}_+^{(1)}, \overline{\mathbf{Z}}_+^{(3)}] = [\overline{\mathbf{Z}}, \overline{\mathbf{Z}}^{(e)}, \overline{\mathbf{Z}}_{w_2}]$. The number n denotes the number of columns in $\overline{\mathbf{Z}}_A^{(\boldsymbol{\ell})}$, and $\mathbf{M}_A^{(\boldsymbol{\ell})} = \mathbf{I}_T - \overline{\mathbf{Z}}_A^{(\boldsymbol{\ell})} \left(\overline{\mathbf{Z}}_A^{(\boldsymbol{\ell})\prime} \overline{\mathbf{Z}}_A^{(\boldsymbol{\ell})}\right)^{\dagger} \overline{\mathbf{Z}}_A^{(\boldsymbol{\ell})\prime}$.

A desirable property of the IC selection procedure is that it identifies the CSA that bring in new information about the factors in $\mathbf{Z}_i$ given what is already present in $\overline{\mathbf{Z}}$. Variables or weights that are uninformative, or informative on factors that do not feature in $\mathbf{Z}_i$ will be excluded (asymptotically). There is, however, no guarantee that the RC will also hold for the chosen CSA unless one is certain that the proposal set contains sufficient informative candidates (see Karabiyik et al., 2019). For example, if the IC does not select additional CSA besides $\overline{\mathbf{Z}}$, this could be either because the rank condition is satisfied with $\overline{\mathbf{Z}}$, or because no further informative CSA are available in the proposal set $\overline{\mathbf{Z}}_+$. Therefore, the IC method alone does not allow one to distinguish between these two scenarios.

To overcome this problem, we propose combining the IC criterion with our proposed RC classifier. Such strategy circumvents the problem of potential failure of the rank condition even when additional CSA have been selected. Our strategy is outlined in Algorithm 1:

Algorithm 1: CCE$_A$ algorithm

```
(1) Estimate the model parameters using the standard CCE approach and calculate $IC_0 = \ln|\Sigma_{i=1}^{N}\mathbf{Z}_i'\mathbf{M}\mathbf{Z}_i/NT| + g(n)$. Proceed to step 2;
(2) Evaluate the rank condition for $\overline{\mathbf{Z}}$. If $\widehat{RC} = 1$, no further steps are required. If $\widehat{RC} = 0$, proceed to step 3;
(3) Employ the IC in Eq. (25) to select from $\overline{\mathbf{Z}}_+ = \{\overline{\mathbf{Z}}_+^{(1)}, \overline{\mathbf{Z}}_+^{(2)}, \overline{\mathbf{Z}}_+^{(3)}, \dots\}$ the set of CSA that are relevant for the factors in $\mathbf{Z}_i$. That is, define $\boldsymbol{\ell}^* = \arg\min_{\boldsymbol{\ell}} IC(\boldsymbol{\ell})$;
(4) If $IC(\boldsymbol{\ell}^*) \leq IC_0$, evaluate the rank condition for $\overline{\mathbf{Z}}_A = [\overline{\mathbf{Z}}, \overline{\mathbf{Z}}_+^{(\boldsymbol{\ell}^*)}]$ and proceed to step 5, else proceed to step 6;
(5) If $\widehat{RC}(\overline{\mathbf{Z}}_A) = 1$, estimate the model with the CCE$_A$ estimator based on $\overline{\mathbf{Z}}_A$. No further steps are required. If $\widehat{RC}(\overline{\mathbf{Z}}_A) = 0$, proceed to step 6;
(6) $\overline{\mathbf{Z}}_+$ does not contain sufficient informative expansions to restore the rank condition in the model. Add new potential expansions to $\overline{\mathbf{Z}}_+$ and return to step 3;
```

Remark 4.1. An alternative approach would be to evaluate the RC for all combinations of potential augmentations until $\widehat{RC} = 1$. However, such approach bares the risk of selecting CSA that load on different factors than those in $\mathbf{Z}_i$, and so they are irrelevant for approximating the factor space. This follows from the easily shown fact that such CSA will increase the rank of the augmented loading matrix, despite being irrelevant, and they will therefore be incorrectly favored by the classifier. One could then falsely conclude that the RC is satisfied with such augmentations. This is especially the case in the absence of truly informative (i.e. relevant) CSA in $\overline{\mathbf{Z}}_+$. A preliminary pass-through by the IC selection eliminates such irrelevant options before they are considered by the classifier. For this reason, it is crucial to combine the RC classifier with the IC pre-selection step, as in Algorithm 1.

5 Monte Carlo Simulation

In this section we investigate the small sample behavior of the rank condition classifier proposed in Section 3 using Monte Carlo simulations.

5.1 Design

Data are generated from Eq. (3), broadly following the design of Westerlund and Urbain (2013). We set $m = 2$, $K = 1$, $\beta = 3$ and sample the time series (columns) in $\mathbf{F}$, ε_i and $\mathbf{V}_i$ assuming independent autoregressive processes with a common AR coefficient $\rho = 0.8$ and normally distributed mean zero innovations with variance $(1 - \rho^2)$ for the factors and $(1-\rho^2)/2$ for the idiosyncratic errors. For the factor loadings γ_i and $\mathbf{\Gamma}_i$, we specify the following three scenarios:

- Experiment 1: $\gamma_i = [3,2]' + \boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i \sim N(\mathbf{0}_2, \mathbf{I}_2)$, and $\mathbf{\Gamma}_i = \gamma_i + [-2,0]'$.
- Experiment 2: $\gamma_i = \begin{cases} [0,2]' + \boldsymbol{\eta}_i & \text{for } i = 1,\dots,\lfloor N/2 \rfloor \\ [2,0]' + \boldsymbol{\eta}_i & \text{for } i = \lfloor N/2 \rfloor + 1,\dots,N \end{cases}$
  with $\boldsymbol{\eta}_i \sim N(\mathbf{0}_2, \mathbf{I}_2)$ and $\mathbf{\Gamma}_i = \gamma_i$.
- Experiment 3: $\gamma_i \sim N(\mathbf{0}_2, \mathbf{I}_2)$ and $\mathbf{\Gamma}_i = \gamma_i$.

This design implies that the rank condition is satisfied in Experiment 1 for the simple CSA $\overline{\mathbf{Z}}$ ($\varrho = m = 2$). Therefore, the standard CCE estimator is consistent. In Experiment 2, the basic CSA contain some information for estimating the factors ($\varrho = 1$), yet not sufficient to satisfy the rank condition and consistently estimate the full factor space ($\varrho = 1 < 2 = m$). Since the loadings in $\mathbf{y}_i$ and $\mathbf{X}_i$ are (perfectly) correlated, the standard CCE estimator is not consistent. In Experiment 3 the standard CSA contain no information at all about the factors ($\varrho = 0 < m$), in which case consistent CCE estimation is also not possible with $\overline{\mathbf{Z}}$.

The purpose of the classifier in Eq.(18) is to identify whether or not the RC holds true within each of the scenarios above. To assess this ability in finite samples, we evaluate the RC in each MC iteration, using Algorithm 1 of Section 4. The number of factors in Eq.(18) is estimated using the Growth Ratio (GR) statistic of Ahn and Horenstein (2013), setting $m_{max} = 7$.[14] The rank of the loading matrix (ϱ) is estimated using the procedure of Robin and Smith (2000) and a random dimension reduction matrix $\mathbf{\Psi}$ with i.i.d. standard-normal entries, and the nominal significance level of the test sequence specified by the function $\alpha_N = c\alpha N^{-1/\gamma}$, with $c = 20$, $\gamma = 1$ and $\alpha = 5\%$. Results for alternative $\mathbf{\Psi}$ (fold-over or sub-sample) are also available upon request.[15]

Additional CSA are constructed from alternative weighting schemes or external covariates:

$$\overline{\mathbf{Z}}_{w,1} = \sum_{i=1}^{N} \mathbf{Z}_i w_{i,1}, \qquad w_{i,1} = \begin{cases} 1/N_1 & \text{for } i = 1,\dots,N/2, \\ 0 & \text{for } i = N/2+1,\dots,N, \end{cases} \tag{26}$$

$$\overline{\mathbf{Z}}_{w,2} = \sum_{i=1}^{N} \mathbf{Z}_i w_{i,2}, \qquad w_{i,2} = \begin{cases} 0 & \text{for } i = 1,\dots,N/2, \\ 1/(N-N_1) & \text{for } i = N/2+1,\dots,N, \end{cases} \tag{27}$$

which results in CSA calculated over respectively the first ($\overline{\mathbf{Z}}_{w,1}$) and second ($\overline{\mathbf{Z}}_{w,2}$) group of $N/2$ cross-sections. This choice of weights presumes the existence of an exogenous grouping of the cross sections, which coincides with Experiment 2, and it is as such an appropriate RC-restoring expansion in this experiment, but not for experiments 1 and 3,

[14] Results based on the "Edge Distribution" estimator of Onatski (2010) are also available upon request.

[15] We also applied the alternative rank estimator by Kleibergen and Paap (2006), but this was found to be less effective than the main approach suggested in this paper. Results are available upon request.

where no such grouping exists.
We also provide candidate CSA originating from the $T \times 2$ matrix of external variables

$$\mathbf{Z}_i^{(e)} = \mathbf{F}\mathbf{C}_i^{(e)} + \boldsymbol{\epsilon}_i^{(e)},$$

with the columns of $\boldsymbol{\epsilon}_i^{(e)}$ generated as an AR(1) process with autoregressive coefficient $\rho = 0.8$ and mean zero normally distributed innovations with variance $(1-\rho^2)/2$, and

$$\mathbf{C}_i^{(e)} = \begin{bmatrix} 2.5 & 1 \\ 1 & 2.5 \end{bmatrix} + \boldsymbol{\eta}_i^{(e)}, \qquad vec(\boldsymbol{\eta}_i^{(e)}) \sim N(\mathbf{0}_4, \mathbf{I}_4).$$

As the $\mathbf{Z}_i^{(e)}$ load on the same factors as those in $\mathbf{Z}_i$, the matrix $\overline{\mathbf{Z}}^{(e)} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{Z}_i^{(e)}$ is an informative, RC-restoring, expansion in experiments 2 and 3. We also accommodate in our simulations the fact that in practice not all external variables will load on the same factors as those in $\mathbf{Z}_i$. These irrelevant candidates are generated from

$$\mathbf{Z}_i^{(g)} = \mathbf{G}\mathbf{C}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)},$$

where the factors $\mathbf{G}$, loadings $\mathbf{C}_i^{(g)}$ and innovations $\boldsymbol{\epsilon}_i^{(g)}$ follow the same DGP as $\mathbf{F}$, $\mathbf{C}_i^{(e)}$ and $\boldsymbol{\epsilon}_i^{(e)}$ but are independently generated from the latter. As such, $\mathbf{Z}_i^{(g)}$ is informative about $\mathbf{G}$ but not $\mathbf{F}$, and $\overline{\mathbf{Z}}^{(g)}$ is therefore not an appropriate expansion in any of the considered experiments. The total set of candidate expansions that is fed into Algorithm 1 is thus a mixture of both relevant and uninformative candidates, and is given by

$$\overline{\mathbf{Z}}_{+} = [\overline{\mathbf{Z}}_{w,1}, \overline{\mathbf{Z}}_{w,2}, \overline{\mathbf{Z}}^{(e)}, \overline{\mathbf{Z}}^{(g)}]. \tag{28}$$

In accordance with Algorithm 1, the augmented estimator CCE_{A} selects expansions from $\overline{\mathbf{Z}}_{+}$ using the Information Criterion by Karabiyik et al. (2019) given in Eq. (25), and the RC is re-evaluated in case expansions have been chosen.

We generate 2000 datasets for each combination of $T = (20, 50, 100, 200)$ and $N = (20, 50, 100, 200, 500, 1000)$ and calculate the under/over-estimation frequencies for $\widehat{\varrho}$ and $\widehat{m}$, and the classification accuracy of the RC classifier $\widehat{RC}$, i.e., the % of Monte Carlo draws where the RC is correctly evaluated. When the RC is not satisfied for the standard CCE estimator (experiments 2 and 3), we also consider the CCE_{A} estimator and compute the 'RC satisfied rate' as the % of Monte Carlo draws where Algorithm 1 selects expansions that restore the rank condition. Note that the classification accuracy of the RC classifier for the CCE_{A} estimator can be split into 'Sensitivity' (i.e. the rate of correct detection that the rank condition holds when the right expansions are selected) and 'Specificity' (i.e. the rate of correct detection that the rank condition does not hold when insufficient expansions are selected). These results are reported in Appendix B, where we also report the specific expansions selected by Algorithm 1 and estimation summaries for β.

5.2 Estimating ϱ and m

We start our discussion with an overview of the performance of the estimators for ϱ and m that we feed into the RC classifier. The results are presented in A/B format in Table 1, with A and B the percentage of MC iterations where ϱ or m are respectively under- and over-estimated. The left hand panel contains results for estimating the rank ϱ of the loading matrix and reveals that both the over-and under-estimation frequencies tend to zero as $N \to \infty$. This confirms our claim that ϱ can be estimated consistently from $\overline{\mathbf{Z}}$. It is clear however, that the rank estimator is nevertheless somewhat sensitive to the size of the cross-section dimension, which needs to be sufficiently large (i.e., N of at least 50) to achieve an accuracy of 75%. In contrast, performance of the rank estimator is largely invariant to the size of T, which supports the projection strategy to guarantee computability of the estimator and large N consistency when also $T \to \infty$. We find that the i.i.d. standard-normal projection $\mathbf{\Psi}$ employed to obtain Table 1 is indeed rank-preserving. Alternative dimension reduction transformations (data omission, averaging) yield similar results and are therefore omitted to save space. Finally, the rank estimator is clearly rather conservative in the sense that the true rank is more likely to be under-estimated than over-estimated. This is by construction, and a consequence of our chosen significance level $\alpha_N = 20\alpha N^{-1}$, of which its fast decay with N implies that strong evidence against the null $\varrho = \varrho^*$ is required before it is rejected in favor of a higher rank $\varrho > \varrho^*$. Yet, the observed under-estimation frequency is reasonable and vanishes sufficiently fast with N. In additional (unreported) simulation results it has been verified that a slower rate of decay on the significance level (say $\alpha_N = 3\alpha N^{-1/3}$) reduces under-estimation but leads to over-estimation of the true rank. We prefer a conservative rank estimator for this classification context as it reduces the risk of falsely concluding that the RC holds when the CCE estimator will in fact be inconsistent.

The right-hand panel of Table 1 reports results for estimating the number of factors $m = 2$ with the Growth Ratio (GR) estimator of Ahn and Horenstein (2013). The estimator performs very well despite the high serial dependence in the generated data, in which case many of its competitors in the literature tend to behave more poorly (results for some alternative estimators are available upon request). In contrast to the rank estimator, which is sensitive to N, the finite sample performance of the GR approach appears to be primarily driven by the time series dimension T. Yet, its small sample performance is more than adequate as the approach displays low error frequencies even when $T = 20$, and identifies m without error when $T > 50$.

Table 1: Under/over-estimation frequency of the estimators for ϱ and m

		$\widehat{\varrho}$				$\widehat{m}$			
	(N,T)	20	50	100	200	20	50	100	200
Experiment 1	20	35/0	26/0	34/0	33/0	14/15	4/0	0/0	0/0
$\varrho = 2, m = 2$	50	23/0	20/0	20/0	19/0	5/7	1/0	0/0	0/0
	100	14/0	20/0	16/0	14/0	4/8	0/0	0/0	0/0
	200	11/0	12/0	11/0	10/0	7/6	0/0	0/0	0/0
	500	9/0	8/0	8/0	8/0	7/5	0/0	0/0	0/0
	1000	5/0	7/0	7/0	5/0	5/6	0/0	0/0	0/0
Experiment 2	20	32/3	33/3	26/4	27/3	10/6	2/0	0/0	0/0
$\varrho = 1, m = 2$	50	19/2	14/2	16/4	17/1	5/3	0/0	0/0	0/0
	100	13/1	4/1	10/0	8/0	8/2	0/0	0/0	0/0
	200	7/1	8/1	8/0	5/0	6/4	0/0	0/0	0/0
	500	5/0	1/0	3/0	3/0	10/1	0/0	0/0	0/0
	1000	2/0	1/0	2/0	3/0	8/1	0/0	0/0	0/0
Experiment 3	20	0/7	0/7	0/8	0/7	14/15	4/0	0/0	0/0
$\varrho = 0, m = 2$	50	0/3	0/3	0/3	0/2	5/7	1/0	0/0	0/0
	100	0/0	0/1	0/1	0/1	4/8	0/0	0/0	0/0
	200	0/2	0/1	0/1	0/1	7/6	0/0	0/0	0/0
	500	0/0	0/0	0/0	0/0	7/5	0/0	0/0	0/0
	1000	0/0	0/0	0/0	0/0	5/6	0/0	0/0	0/0

Notes: (i) Based on 2 000 MC iterations. (ii) Reported in the left hand panel is the percentage of under/over- estimation of the true rank ϱ by the rank estimator $\widehat{\varrho}$ applied to $\overline{\mathbf{Z}}$, with $\mathbf{\Psi}$ drawn from the standard-normal distribution, $\alpha_N = 20\alpha N^{-1}$ and $\alpha = 5\%$. (iii) The right hand panel is the percentage of under/over estimation of the true number of factors $m = 2$ by the Growth Ratio estimator with $m_{max} = 7$.

5.3 Evaluating the rank condition

Experiment 1: rank condition satisfied

We start with Experiment 1, in which the rank condition is satisfied for the CCE estimator that uses the standard set of CSA in $\overline{\mathbf{Z}}$. The classification accuracy reported in Table 2 shows that the $\widehat{RC}$ classifier is reasonably accurate in detecting that the rank condition is indeed satisfied. Even for smaller samples, the RC is correctly confirmed for at least 70% of the MC iterations, the only exception being the smallest $N = 20$ setting where the lowest rate is 59%. As the sample size grows, the accuracy improves and we find that it tends to 1 as both $(N, T) \to \infty$, as required. The results also show that the main determinant for finite sample performance is the cross-section dimension N, rather than T. This is as expected from the results in Table 1, which show that $\widehat{\varrho}$ is more prone to finite sample error than is $\widehat{m}$, which is practically error-less when $T \geq 50$, and $\widehat{\varrho}$ furthermore converges at a slower rate and only as N grows. Hence, $\widehat{\varrho}$ is the main driver of the finite sample performance of $\widehat{RC}$, and therefore, in line with the properties of the CCE estimator itself, it will mainly be N that needs to be sufficiently large to be able to correctly assess the rank condition in practice. Yet, it is clear that N does not need to be very large to obtain good classification accuracy, nor does the conservative specification of $\widehat{\varrho}$ lead to too many false conclusions. Note furthermore that the samples where we incorrectly obtain $\widehat{RC} = 0$ classifications for the CCE estimator prompted the application of the augmentation strategy outlined in Algorithm 1 of section 4. As shown in Table 8 in the appendix, an expansion was only selected in the smallest samples and in at most 2% of the MC iterations. Hence, the rank evaluation results for the augmented CCE$_\text{A}$ estimator reported in the right panel of Table 2 are almost identical to those for the CCE estimator.

Table 2: Evaluating the rank condition: Experiment 1

		CCE				CCE$_\text{A}$			
	(N, T)	20	50	100	200	20	50	100	200
Classification accuracy	20	0.59	0.73	0.66	0.66	0.64	0.73	0.66	0.66
	50	0.71	0.80	0.79	0.80	0.73	0.80	0.79	0.80
	100	0.80	0.80	0.84	0.86	0.82	0.80	0.84	0.86
	200	0.84	0.88	0.89	0.90	0.84	0.88	0.89	0.90
	500	0.86	0.92	0.91	0.91	0.86	0.92	0.91	0.91
	1000	0.89	0.92	0.93	0.95	0.89	0.92	0.93	0.95

Notes: (i) Based on 2 000 MC iterations. (ii) Reported is the Classification Accuracy (CA), which is the proportion of MC samples in which the classifier $\widehat{RC}$ defined in Eq. (18) correctly identifies whether the RC is satisfied or not. (iii) The RC classifier uses the GR estimator of Ahn and Horenstein (2013) with $m_{max} = 7$ to estimate m, and the Robin and Smith (2000) rank estimator with a standard-normal projection matrix and significance level $\alpha_N = 20\alpha N^{-1}$ to estimate ϱ. (iv) The left panel evaluates the rank condition for the standard CCE estimator that uses the matrix of CSA $\overline{\mathbf{Z}}$ to control for the unobserved factors. The right panel evaluates the rank condition for the CCE$_\text{A}$ estimator, which is the outcome of Algorithm 1 presented in section 4. That is, if $\widehat{RC} = 1$ for $\overline{\mathbf{Z}}$, then only $\overline{\mathbf{Z}}$ is employed in the estimation. If on the other hand $\overline{\mathbf{Z}}$ yields $\widehat{RC} = 0$, then expansion CSA are selected from $\overline{\mathbf{Z}}_+$ using the IC in (25). The rank condition is then re-evaluated for the augmented set of CSA.

Experiment 2: rank condition violated for basic weights

Next, we discuss experiment 2, where the rank condition is violated when using the standard set of CSA $\overline{\mathbf{Z}}$. As the factor loadings are (perfectly) correlated, the CCE estimator is inconsistent for β in this setting (this can also be seen from the estimation results in Table 11 in Appendix B). The left 'CCE' panel of the classification results in Table 3 shows that the RC-classifier strongly signals that the rank condition is violated for the CCE estimator. The proportion of samples where the classifier wrongly concludes that the RC holds quickly diminishes as $(N, T) \to \infty$, with the classification accuracy amounting to more than 90% for most sample sizes.

When the rank condition is found to be violated, Algorithm 1 is applied to resolve the rank deficiency of the CCE estimator by letting the IC search among the proposal expansions for additional CSA. In this experiment this results in the selection of at least one of the valid augmentations $(\overline{\mathbf{Z}}_{w,1}, \overline{\mathbf{Z}}_{w,2}, \overline{\mathbf{Z}}^{(e)})$ in the majority of combinations of N and T (see Table 8 in Appendix B). Accordingly, the rank condition was successfully restored in 91% of the MC iterations even when $T = N = 20$. This confirms the effectiveness of the IC by Karabiyik et al. (2019) for selecting appropriate expansions. The proportion of samples where the RC is restored is given in the lower panel for the CCE_A estimator, and can be seen to converge to 1 as $(N, T) \to \infty$. Hence, Algorithm 1 leads to a consistent CCE_A estimator as $(N, T) \to \infty$ (when provided with appropriate rank-increasing CSA), which is also confirmed by the estimation results for β in Table 11 of Appendix B. In addition, note that the algorithm also performs well in finite samples with a high success rate. The cases where the RC is not satisfied for CCE_A are due to the miss-classification as $\widehat{RC} = 1$ in the 'CCE' panel, which indeed vanishes as the sample size grows.

In practice, selecting expansion CSA with the IC does not guarantee that the rank condition is also satisfied, leaving the researcher unsure about the state of the RC. Hence, Algorithm 1 incorporates a re-evaluation with the classifier after expansions have been chosen. The top right panel of Table 3 reveals that this re-evaluation is able to confirm with good accuracy that the rank condition is satisfied in those cases where the right expansions have been selected (see also the Sensitivity, or the rate of correct detection for the RC holds cases, given in Table 9 of Appendix B). The overall classification accuracy is over 70% in the smallest samples and gradually converges to 1 as $(N, T) \to \infty$. The few cases where the RC remains violated are all due to an incorrect $\widehat{RC} = 1$ conclusion for $\overline{\mathbf{Z}}$ in the first step, which does not prompt action by the algorithm. These RC failures remain undetected by Algorithm 1 and hence lead to a Specificity ('rate of correct detection of RC fail cases') of technically 0%. Note, however, that this amounts to max 8% of the MC samples even for $N = T = 20$.

Table 3: Evaluating the rank condition: Experiment 2

		CCE				CCE_A			
	(N, T)	20	50	100	200	20	50	100	200
Classification accuracy	20	0.92	0.94	0.95	0.97	0.73	0.89	0.93	0.95
	50	0.93	0.97	0.96	0.98	0.86	0.93	0.95	0.98
	100	0.92	0.99	1.00	1.00	0.89	0.99	1.00	1.00
	200	0.94	0.98	1.00	0.99	0.90	0.98	1.00	0.99
	500	0.90	0.99	1.00	1.00	0.89	0.99	1.00	1.00
	1000	0.91	1.00	1.00	1.00	0.90	0.99	1.00	1.00
RC satisfied rate	20	Always 0 (by construction)				0.91	0.94	0.95	0.97
	50					0.93	0.97	0.96	0.98
	100					0.92	0.99	1.00	1.00
	200					0.94	0.98	1.00	0.99
	500					0.90	0.99	1.00	1.00
	1000					0.91	1.00	1.00	1.00

See notes to Table 2. The 'RC satisfied rate' is the % of MC samples in which the algorithm behind CCE_A selects CSA augmentations that restore the rank condition.

Experiment 3: rank condition violated

Experiment 3 is a setting where the loading matrix $\mathbf{C}$ for the standard CSA is rank zero. Intuitively, the effect of the factors is averaged out in $\overline{\mathbf{Z}}$ such that the CSA are uninformative for estimating the factor space and the CCE estimator is inconsistent (see Table 12 in Appendix B for confirmation). Table 4 summarizes the ability of our evaluation procedure to detect this case where the RC is ′severely′ violated. The top panel reveals that our method is highly accurate in this setting even for very small N. This is due to the large discrepancy between $m = 2$ and $\varrho = 0$, the latter of which would need to be overestimated by 2 in order to incorrectly conclude that the RC is satisfied. As we have specified a conservative estimator for ϱ, such an over-estimation almost never occurred (recall the bottom panel of Table 1).

Given the strong signal by the classifier that the RC is violated, Algorithm 1 in the ′CCE$_A$′ panel has led to a search for expansion CSA in nearly all MC samples. We find that the sole rank-restoring expansion $\overline{\mathbf{Z}}^{(e)}$ was selected with high probability, as indicated by the high proportion of samples for which the RC has been restored (see the bottom panel of Table 4). With the exception of $T = 20$ samples, the irrelevant expansions ($\overline{\mathbf{Z}}_{w,1}, \overline{\mathbf{Z}}_{w,2}, \overline{\mathbf{Z}}^{(g)}$) are successfully excluded, which confirms the good properties of the IC also in this experiment. This can be seen from Table 8 in Appendix B. Note, however, that compared to Experiment 2 the classifier appears less capable to confirm that the rank condition is restored when N is very small. Accuracy is only 40% when $N = 20$ (see also the low Sensitivity in Table 9 in Appendix B). Closer analysis reveals that this is caused by a relatively large under-estimation rate (60%) of the true rank in $N = 20$ samples when the correct expansion was selected. A possible cause is that the expanded matrix $\overline{\mathbf{Z}}_A = [\overline{\mathbf{Z}}, \overline{\mathbf{Z}}^{(e)}]$ has a potential rank (number of columns=4) which is twice the true rank (2). This suggests a relatively high level of estimation noise, and a cross-section dimension of $N = 20$ appears too small to estimate the rank well in such cases. Yet, the performance of the estimator improves quickly with N and classification accuracy recovers to 85% or higher for $N = 100$. This confirms our earlier conclusion that sufficiently large N is key for good evaluation of the rank condition.

As a final experiment, we consider also the empirically relevant scenario where the proposal set $\overline{\mathbf{Z}}_+$ does not contain sufficient informative CSA to restore the rank condition. To that end, we report in the CCE$_{A,sub}$ panel of Table 4 the outcomes of Algorithm 1 when the set of proposal expansions is $\overline{\mathbf{Z}}_{+,sub} = \{\overline{\mathbf{Z}}_{w,1}, \overline{\mathbf{Z}}_{w,2}, \overline{\mathbf{Z}}^{(g)}\}$ in stead of $\overline{\mathbf{Z}}_+$. Hence, $\overline{\mathbf{Z}}_{+,sub}$ contains insufficient valid expansions to restore the RC, and Algorithm 1 should signal that the RC remains violated even when expansions have been selected from it. It is furthermore important that the IC does not select $\overline{\mathbf{Z}}^{(g)}$, the CSA that loads on factors not in $\mathbf{Z}_i$, as it would lead to false conclusions that the RC is satisfied by the classifier (see

Remark 4.1). This makes the setting particularly challenging. The results summarized in the CCE$_{\mathrm{A},sub}$ panel of Table 4 show, however, that the classifier confirms with high accuracy that the RC fails even after expansions were chosen. This is also visible from the Sensitivity/Specificity breakdown in Table 9 of Appendix B. In dept analysis reveals that the few miss-classified cases are primarily caused by selection of the invalid $\overline{\mathbf{Z}}^{(g)}$ expansion by the IC, which has indeed prompted false conclusions that $\widehat{RC} = 1$. Yet, since the IC is able to successfully exclude this expansion when T is sufficiently large ($T > 20$ suffices in this experiment), miss-classification indeed vanishes as N and T grow, as required. Unreported simulation results confirm that alternatives to Algorithm 1 which do not include a pre-selection step by the IC lead to erroneous conclusions that the RC can be restored with $\overline{\mathbf{Z}}_{+,sub}$.

Table 4: Evaluating the rank condition: Experiment 3

		CCE				CCE$_{\mathrm{A}}$				CCE$_{\mathrm{A},sub}$			
	(N,T)	20	50	100	200	20	50	100	200	20	50	100	200
Classification accuracy	20	0.98	0.99	0.99	1.00	0.38	0.41	0.38	0.40	0.92	0.96	0.92	0.94
	50	0.99	0.99	0.99	1.00	0.65	0.68	0.67	0.73	0.97	0.97	0.98	0.99
	100	1.00	1.00	1.00	1.00	0.86	0.94	0.96	0.91	0.98	0.99	0.99	1.00
	200	0.99	1.00	1.00	1.00	0.91	0.98	0.98	0.96	0.97	0.99	0.99	0.98
	500	1.00	1.00	1.00	1.00	0.93	0.99	0.99	0.97	0.96	0.99	1.00	1.00
	1000	1.00	1.00	1.00	1.00	0.93	0.99	0.99	0.99	0.98	1.00	0.99	0.99
RC satisfied rate	20	Always 0 (by construction)				0.92	0.97	0.99	0.99	Always 0 (by construction)			
	50					0.96	0.99	0.99	1.00				
	100					0.96	0.99	1.00	1.00				
	200					0.96	1.00	1.00	1.00				
	500					0.98	0.99	1.00	1.00				
	1000					0.98	1.00	1.00	1.00				

See notes to Tables 2 and 3. CCE$_{\mathrm{A},sub}$ refers to using Algorithm 1, with the set of potential augmentations given by $\overline{\mathbf{Z}}_{+,sub} = \{\overline{\mathbf{Z}}_{w,1}, \overline{\mathbf{Z}}_{w,2}, \overline{\mathbf{Z}}^{(g)}\}$ instead of $\overline{\mathbf{Z}}_{+}$.

6 On the impact of the Dodd-Frank Act on the profitability of U.S. banks

Banks play an important role in the functioning of national economies because they act as financial intermediaries between savers and borrowers, and facilitate the pricing and allocation of risks. Studies on the profitability of banking institutions are vital for obtaining better understanding of the causes of financial crises, economic recessions and growth. On the one hand, profits constitute the first line of defense against losses from credit impairment, since retained earnings are an important source of capital. On the other hand, when it comes to large banks, high profitability may also signal excessive market power through stronger brand image or implicit regulatory protection; this is the so-called "too-big-to-fail" (TBTF) hypothesis, which postulates that large financial institutions may be so widely interconnected to the rest of the economy that their failure would generate a disastrous domino effect for the whole economy. Thus, to the extent that governments effectively subsidize downsize risk for financial institutions with TBTF status, large banks face artificially lower costs of capital, and thus reap more profits (see e.g., Cetorelli and Traina, 2018).

There is a large number of studies that analyse drivers of bank profits in general (see e.g., Athanasoglou et al., 2008; Baker and Wurgler, 2015; Goddard et al., 2011, 2004; Iannotta et al., 2007; Lee and Hsieh, 2013; Staikouras and Wood, 2004). There is also a fairly substantial literature focusing on the TBTF hypothesis (see e.g. Gropp and Vesala, 2004; Hakenes and Schnabel, 2011; Morgan and Stiroh, 2005; Sironi, 2003; Stern and Feldman, 2009; Völz and Wedow, 2011). The bulk of this literature provides evidence that government bailout guarantees may distort market discipline, inducing excessive risk-taking and morally hazardous behavior (Mattana et al., 2015).

The present illustration contributes to this literature by examining the impact of the well-known "Dodd-Frank Act" (DFA) on profitability in the U.S. banking sector. The DFA is a U.S. federal law enacted during 2010, aiming "to promote the financial stability of the United States by improving accountability and transparency in the financial system, to end "too big to fail", to protect the American taxpayer by ending bailouts, to protect consumers from abusive financial services practices, and for other purposes".[16] In a nutshell, the DFA has instituted a new failure-resolution regime, which seeks to ensure that losses resulting from bad decisions by managers are absorbed by equity and debt holders, thus potentially reducing moral hazard.

Existing empirical evidence on the extent to which the DFA has alleviated the TBTF is relatively sparse and not in agreement. For example, while Baily et al. (2020) conclude on a positive influence of the DFA towards resolving moral hazard, other studies point in the opposite direction (see e.g. Bordo and Duca, 2018). In what follows, we apply the CCE estimator and rank test methodology developed in the present paper to shed some light on this important topic.

6.1 Model Specification

We make use of a panel data set consisting of 450 U.S. banking institutions, each one observed over 56 quarters. The sample spans the period 2006:Q1–2019:Q4 and includes the financial crisis (2007–2009).[17] We analyse the impact of major drivers of bank profitability, with emphasis on bank size. Thus, we specify the following model:

[16] See https://www.cftc.gov/sites/default/files/idc/groups/public/@swaps/documents/file/hr4173_enrolledbill.pdf.

[17] All data are publicly available and they have been downloaded from the Federal Deposit Insurance Corporation (FDIC) website. See https://www.fdic.gov/.

$$ROA_{it} = \beta_1^{(\ell)} SIZE_{it} + \beta_2^{(\ell)} CAR_{it} + \beta_3^{(\ell)} LIQUIDITY_{it} + \beta_4^{(\ell)} QUALITY_{it} + \beta_5^{(\ell)} NPL_{it} + u_{it};$$

$$u_{it} = \eta_i + \boldsymbol{\lambda}_i' \mathbf{f}_t + \varepsilon_{it}, \tag{29}$$

where $i = 1, \ldots, N(= 450)$, $t = 1, \ldots, T(= 56)$, $\ell = \tau_1 \mathbb{1}\{t < \tau\} + \tau_2 \mathbb{1}\{t \geq \tau\}$ and τ signifies the first quarter of the first full year after which the DFA became effective, i.e. 2011:Q1. Essentially, the model above is estimated for two different sub-periods, namely 2006:Q1–2010:Q4 and 2011:Q1–2019:Q4; the first sub-period belongs to the Basel I-II period, whereas the second one corresponds to the DFA and coincides with the introduction of the Basel III internationally.[18]

The variables of the model are defined as follows:

- ROA_{it} denotes the return on assets, defined as annualized net income expressed as a percentage of average total assets on a consolidated basis;
- $SIZE_{it}$ is proxied by the natural logarithm of bank total assets;
- CAR_{it} stands for "capital adequacy ratio", which is proxied by the ratio of Tier 1 (core) capital over average total assets minus ineligible intangibles. Higher values of this ratio imply higher levels of capitalisation;
- $LIQUIDITY_{it}$ is given by the loan-to-deposit (LTD) ratio. A higher value of this ratio implies a lower level of liquidity;
- $QUALITY_{it}$ represents the quality of bank assets and is computed as the total amount of loan loss provisions (LLP) expressed as a percentage of assets. Thus, a higher level of loan loss provisions indicates lower quality;
- NPL_{it} is a measure of risk, and denotes the ratio of non-performing loans to total loans for bank i at time period t. Higher values of the NPL ratio indicate that banks ex-ante took higher lending risk and therefore they have accumulated ex-post more bad loans;

The error term u_{it} is composite. In particular, η_i captures bank-specific effects, such as ownership and location, both of which can be important factors for profitability (Zimmerman, 1996). The $m \times 1$ vector $\mathbf{f}_t$ denotes unobserved economy-wide factors that influence bank profits, albeit with heterogeneous intensities (absorbed by the bank-specific factor loadings), $\boldsymbol{\lambda}_i$. Last, ε_{it} is a purely idiosyncratic error.

The above set of explanatory variables originate from bank accounts (balance sheets and/or profit and loss accounts) and are tied to management decisions. As such, they

[18] Basel III is an international regulatory framework for capital standards, which incorporates a set of reforms within the banking sector, designed to improve the regulation, supervision and risk management. In short, largely in response to the credit crisis, the Basel III requires banks to maintain proper leverage ratios and meet certain minimum capital requirements.

are viewed as "internal". Bank profitability is also driven by "external" factors that lie beyond the control of management, such as business cycle effects, monetary shocks and financial innovation. These are absorbed in our model by the common factor component specified in the error term, $\boldsymbol{\lambda}_i' \mathbf{f}_t$. Although in some cases external drivers can be measured and included directly in the model, often the details of measurement may be difficult and/or contentious.[19]

We note that internal and external drivers of bank profitability are likely to be mutually correlated. For example, asset quality may depend on the position of the business cycle, since contractionary phases are typically associated with a higher level of default risk. Therefore, standard panel data approaches that fail to control for external drivers are likely to face an endogeneity issue and, hence, to yield inconsistent parameter estimates. The CCE approach allows for consistent estimation, provided that the rank condition is satisfied such that the external drivers are adequately controlled for.

Some discussion on the interpretation of the parameters that characterize Eq. (29) is noteworthy. To begin with, $\beta_1^{(\ell)}$ reflects the impact of bank size on profits. Thus, $\beta_1^{(\ell)}$ captures the effect of market power and implicit regulatory protection via TBTF. Moreover, β_1^{ℓ} also absorbs the effect of economies of scale. Such scale effects will be positive (negative) if there exist economies (diseconomies) of scale. However, to the extent that the degree of returns to scale in the banking sector has remained unaltered during the sampling period of the analysis, the difference in the coefficient of $SIZE_{it}$ between the two sub-periods, i.e., $\beta_1^{(\tau_2)} - \beta_1^{(\tau_1)}$, will measure the impact of the Dodd-Frank Act on TBTF, conditional on the remaining covariates.

$\beta_2^{(\ell)}$ and $\beta_3^{(\ell)}$ measure the effect of capital adequacy and liquidity, respectively, on bank profits. In theory, an excessively high CAR could signify that a bank is operating overcautiously and is ignoring potentially profitable investment opportunities. Similarly, a bank that holds a relatively high proportion of liquid assets (hence, $LIQUIDITY$ is relatively low) is unlikely to earn high profits. Therefore, $\beta_2^{(\ell)}$ and $\beta_3^{(\ell)}$ are expected to be negative and positive, respectively.

Poor asset quality can be a major cause of decreased profitability, since higher loan loss provisions tend to impair bank balance sheets by construction. This negative effect is absorbed by $\beta_4^{(\ell)}$.

Finally, $\beta_5^{(\ell)}$ measures the effect of portfolio risk on profits. The main idea is that when banks are less exposed to risk, shareholders might be willing to accept a lower return on assets. If this is true, $\beta_5^{(\ell)}$ would be negative.

For notational convenience of the analysis below, let $\mathbf{Z}_i$ denote the $T \times 6$ matrix with the

[19] For example, how does one measure monetary shocks? Does one look at interest rates or monetary aggregates? Which monetary aggregates? Similarly, how does one proxy financial innovation? For instance, how does one measure embedded leverage in new financial instruments?

observables of the model

$$\mathbf{Z}_i = \left[\mathbf{y}_i, \mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(5)} \right], \tag{30}$$

where $\mathbf{y}_i = [y_{i1}, \ldots, y_{iT}]'$ is a $T \times 1$ vector such that $y_{it} \equiv ROA_{it}$, and similarly for the remaining variables, where $\mathbf{x}_i^{(k)}$ denotes the covariate with coefficient $\beta_k^{(\ell)}$, for $k = 1, \ldots, K (= 5)$.

6.2 Evaluating the RC

Before looking at the results obtained from the CCE approach, it is important to test whether the RC is satisfied. The number of factors m is estimated from the $T \times (K+1)N$ matrix $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_N]$, using the Growth Ratio statistic of Ahn and Horenstein (2013), as explained in Section 3.2. The rank of the matrices of CSA that we consider, to be defined shortly, is determined based on the sequential testing procedure of Robin and Smith (2000), which is described in Section 3.1. Since T is small in both sub-periods of the sample, there is no need to reduce the row-dimensionality using a projection matrix. Thus, we perform the testing procedure on the original CSA matrix.

Standard CCE estimator

We start with the standard CCE estimator of Pesaran (2006), which is based on unweighted CSA of the observables

$$\overline{\mathbf{Z}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_i. \tag{31}$$

We focus initially on the bottom panel of Table 5, which reports results for the evaluation of the rank condition. The first and second columns correspond to the standard CCE estimator applied to the periods 2006:Q1–2010:Q4 (Basel I-II) and 2011:Q1–2019:Q4 (Dodd-Frank Act), respectively.

For the period under Basel I-II we estimate three factors, $\widehat{m} = 3$. The standard set of CSA $\overline{\mathbf{Z}}$ appears to be unable to proxy these factors as the RC is found to be violated, $\widehat{RC} = 0$, for the standard CCE estimator. For the period under the Dodd-Frank Act, we obtain $\widehat{m} = 2$ and the rank condition now appears to hold for the CCE estimator.

Table 5: US bank profitability: Evaluating the rank condition

	CCE		CCE_A	
	Basel I-II	**Dodd-Frank Act**	**Basel I-II**	**Dodd-Frank Act**
$\widehat{m}$	3	2	3	2
$\widehat{\varrho}$	1	3	3	2
$\widehat{RC}$	0	1	1	1

Notes: The columns under Basel I-II report results for the first subsample, which spans 2006:Q1–2010:Q4. The columns under Dodd-Frank Act correspond to the period 2011:Q1–2019:Q4. $\widehat{m}$ is the number of factors estimated from the $T \times (K+1)N$ matrix $\mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_N]$, using the Growth Ratio statistic of Ahn and Horenstein (2013), and $\widehat{\varrho}$ is the rank estimator of Robin and Smith (2000), with $\alpha_N = 20\alpha N^{-1}$ and $\mathbf{\Psi} = \mathbf{I}_T$, as explained in Section 3.2. $\widehat{RC}$ is classifier for the rank condition defined in Eq. (18).

Augmented CCE estimator

Given that the RC is violated for the standard CCE approach in the first sub-period of the sample, we consider a set of potential expansion CSA, which is given by

$$\overline{\mathbf{Z}}_+ = \{\overline{\mathbf{Z}}_+^{(1)}, \overline{\mathbf{Z}}_+^{(2)}, \overline{\mathbf{Z}}_+^{(3)}, \overline{\mathbf{Z}}_+^{(4)}\}. \tag{32}$$

$\overline{\mathbf{Z}}_+^{(1)} \equiv [\overline{\mathbf{x}}^{(6)}, \overline{\mathbf{x}}^{(7)}]$ is a $T \times 2$ matrix, where $\overline{\mathbf{x}}^{(6)}$ and $\overline{\mathbf{x}}^{(7)}$ denote the simple CSA of two external variables, namely the return to equity (ROE), and the tier 1 risk-based capital ratio. ROE is defined as annualized net income expressed as a percent of average total equity on a consolidated basis. The risk-based capital ratio is defined as the tier 1 (core) capital expressed as a percent of risk-weighted assets. The rationale behind using these variables as factor proxies lies in that they present alternative measures of profitability ($\mathbf{y}_i$) and capitalization ($\mathbf{x}_i^{(2)}$), respectively. As such, they are expected to be driven by the same common factors as those entering into the regression model.

$\overline{\mathbf{Z}}_+^{(2)}$ and $\overline{\mathbf{Z}}_+^{(3)}$ denote $T \times (K+1)$ matrices of *weighted* CSA, computed from $\mathbf{Z}_i$ in Eq. (30). $\overline{\mathbf{Z}}_+^{(2)}$ is calculated using as aggregation weight the initial level of bank-specific debt ratio value, which is defined as total liabilities over total assets. This variable has been employed in the literature as a measure of interconnectedness of banks (see Fernandez (2011)). Thus, banks with similar levels of debt ratio may be hit by common shocks in an alike manner and therefore they take a similar weight in the computation of the CSA of $\mathbf{Z}_i$. $\overline{\mathbf{Z}}_+^{(3)}$ uses the size of each bank in the beginning of the sample as averaging weight. This implies that banks of similar size get a similar weight in the computation of $\overline{\mathbf{Z}}_+^{(3)}$.

Finally, $\overline{\mathbf{Z}}_+^{(4)}$ denotes a $T \times 2(K+1)$ matrix of CSA, obtained using two weights that are constructed by grouping banks according to their size. In particular, we take the CSA of the $(K+1)$ observables over the bottom and as well as of the top quintile (i.e., 20%) of banks[20] by using the following two weights:

$$\mathbf{w}_i = \begin{cases} [1/(0.2N), 0]' & \text{if bank } i \text{ is in the bottom quintile,} \\ [0, 0]' & \text{if bank } i \text{ is not in the bottom or top quintile} \\ [0, 1/(0.2N)]' & \text{if bank } i \text{ is in the top quintile} \end{cases}$$

Table 6 reports IC results for each of the suggested additional CSA. Even though the rank condition for the standard CCE estimator was found to be satisfied in the second sub-period, for completeness, we report results in terms of the IC for both sub-periods.

Table 6: US bank profitability: IC for additional CSA

	Basel I-II	**Dodd-Frank Act**
	IC	*IC*
$\overline{\mathbf{Z}}$	-4.149	-7.664
$[\overline{\mathbf{Z}}, \overline{\mathbf{Z}}_+^{(1)}]$	-4.260	-5.510
$[\overline{\mathbf{Z}}, \overline{\mathbf{Z}}_+^{(2)}]$	0.588	-0.516
$[\overline{\mathbf{Z}}, \overline{\mathbf{Z}}_+^{(3)}]$	2.045	0.012
$[\overline{\mathbf{Z}}, \overline{\mathbf{Z}}_+^{(4)}]$	1.529	7.044

Note: the IC criterion is specified in Eq. (25).

Clearly, the unweighted CSA of the two external variables $\mathbf{x}^{(6)}$ and $\mathbf{x}^{(7)}$ included in $\overline{\mathbf{Z}}_+^{(1)}$ appear to provide the most new information, both under Basel I-II, as well as the DFA period. $\overline{\mathbf{Z}}_+^{(1)}$ is, however, not selected as an expansion in the DFA period as the IC calculated without expansions (i.e., with $\overline{\mathbf{Z}}$ only) is $IC_0 = -7.664$. The IC thus confirms the finding in Table 5: no expansions are required in the DFA period as the rank condition is already satisfied. Under Basel I-II, on the other hand, the RC was found to be violated and the IC selects only $\overline{\mathbf{Z}}_+^{(1)}$ as an expansion (the IC based on $\overline{\mathbf{Z}}$ alone is $IC_0 = -4.149$). This is consistent with the argument above that $\overline{\mathbf{Z}}_+^{(1)}$ loads on the same factors as in the regression model, and indicates that the remaining expansions ($\overline{\mathbf{Z}}_+^{(2)}, \overline{\mathbf{Z}}_+^{(3)}, \overline{\mathbf{Z}}_+^{(4)}$) do not provide new information about the factor space given that already present in $\overline{\mathbf{Z}}$, or that they load on different factors than those in $\mathbf{Z}_i$. Therefore, we consider the augmented CCE estimator with the following matrix of CSA:

$$\overline{\mathbf{Z}}_A = [\overline{\mathbf{Z}}, \overline{\mathbf{Z}}_+^{(1)}]. \tag{33}$$

Whether this augmented set of CSA is sufficient to restore the rank condition needs to be verified with the RC classifier. If the RC is still violated, alternative potential expansions

[20] Note that if weights were constructed by splitting *all* banks in two groups, these weights would result in perfect multicollinearity, since we already include simple CSA in the model.

need to be sought (as the IC indicate that only $\overline{\mathbf{Z}}_{+}^{(1)}$ in $\overline{\mathbf{Z}}_{+}$ is relevant). The results for evaluating the RC for this augmented-CCE estimator, denoted as CCE$_{\mathrm{A}}$, are reported in the right panel of Table 5. As we can see, the augmentation has restored the rank condition ($\widehat{RC} = 1$) for the first sub-period. Hence, there is no need to look for further expansions. As expected, the RC remains satisfied in the second sub-period should we also augment the CCE estimator with $\overline{\mathbf{Z}}_{+}^{(1)}$ in that sample.

6.3 CCE and CCE$_{\mathrm{A}}$ estimation results

Table 7 reports CCE and CCE$_{\mathrm{A}}$ estimates of the model parameters for the two sub-periods 2006:Q1–2010:Q4 and 2011:Q1–2019:Q4. For each coefficient, the top row refers to the point estimate and the bottom row refers to the standard error, computed using the parametric sandwich-type formula in Pesaran (2006).

Table 7: US bank profitability: CCE and CCE$_{\mathrm{A}}$ estimation results

	Basel I-II		**Dodd-Frank Act**	
	CCE	CCE$_{\mathrm{A}}$	CCE	CCE$_{\mathrm{A}}$
$\widehat{\beta}_1$ (size)	0.959***	0.647***	0.267*	0.331**
	(0.325)	(0.196)	(0.149)	(0.156)
$\widehat{\beta}_2$ (CAR)	-0.035**	-0.038***	-0.027	-0.026
	(0.017)	(0.015)	(0.021)	(0.021)
$\widehat{\beta}_3$ (liquidity)	1.045***	0.646***	0.964***	0.871***
	(0.364)	(0.251)	(0.170)	(0.192)
$\widehat{\beta}_4$ (quality)	-0.943***	-0.914***	-0.890***	-0.905***
	(0.061)	(0.040)	(0.050)	(0.048)
$\widehat{\beta}_5$ (NPL)	0.016	0.017	-0.027***	-0.025**
	(0.012)	(0.011)	(0.009)	(0.010)

Notes: The columns under Basel I-II report results for the first subsample, which spans 2006:Q1–2010:Q4. The columns under Dodd-Frank Act correspond to the period 2011:Q1–2019:Q4. Standard errors, computed based on the parametric sandwich-type formula in Eq. (74) of Pesaran (2006), are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The above RC evaluation implies that in the first sub-period the CCE$_{\mathrm{A}}$ estimator is consistent, whereas CCE is not. Such discrepancy is mainly noticeable in the estimated coefficients of $SIZE$ and $LIQUIDITY$. In both cases, CCE appears to overestimate the impact of these variables on bank profitability. This provides further evidence that the RC is violated under Basel I-II for CCE.

For the period 2011:Q1–2019:Q4, which corresponds to the Dodd-Frank Act, the RC condition appears to hold for both CCE and CCE$_{\mathrm{A}}$. Thus, in this case it appears that there is no need to augment the model with additional CSA. Notably, the estimated coefficients obtained by the two estimators are not statistically different.

Interestingly, the standard errors for the CCE$_{\mathrm{A}}$ estimator are lower than those of the CCE estimator in the first but not in the second sub-period. This further suggests that the additional CSA do capture nuisance factors that are left in the error term of the CCE estimator in the 2006:Q1–2010:Q4 sub-period, while being irrelevant and hence driving up parameter uncertainty over the 2011:Q1–2019:Q4 sub-period.

Turning to a comparison of the results across the two sub-periods, $SIZE$ appears to be substantially less important in terms of driving profitability of banks under the DFA period compared to the Basel I-II. More specifically, the difference between $\widehat{\beta}_1^{(\tau_1)} = 0.647$ (using the CCE$_{\mathrm{A}}$ estimate) and $\widehat{\beta}_1^{(\tau_2)} = 0.267$ (using the CCE estimate) equals 0.38 and is statistically significant at the 10% level of significance, with a p-value that is roughly equal to 0.061 (one-tailed test).[21]

That is, if large banks exercised market power and implicitly relied on regulatory protection based on a “too-big-to-fail” presumption, such type of behavior seems to be less prevalent after the introduction of the Dodd-Frank Act. This outcome is consistent with the findings of Gao et al. (2018), Cui et al. (2020) and Zhu et al. (2020) and provides evidence that the regulatory reforms introduced by the DFA have succeeded in influencing banks’ behavior in a substantial manner. Further, note that if we use the standard CCE estimator under both sub-periods of the sample, the difference between $\widehat{\beta}_1^{(\tau_1)}$ and $\widehat{\beta}_1^{(\tau_2)}$ amounts to $0.959 - 0.267 = 0.692$. Hence, the impact of the DFA is estimated to be twice as large as that obtained based on our approach. This further highlights the importance of evaluating the rank condition for CCE-type estimators.

Other major differences across the two sub-periods include: (i) profitability is negatively linked with NPL under the DFA, whereas no such link appears to be established under Basel I-II.[22]; (ii) although the coefficient of CAR remains negative under DFA, the effect of capitalisation – conditional on liquidity – is no longer statistically significant at the 10% level.

7 Conclusion

It is well known that the so-called Rank Condition - the requirement that there are at least as many observables containing independent information about the unobserved factors as there are factors in the model - is crucial for the statistical properties of the

[21] The t-statistic is calculated as $t = (0.647 - 0.267)/\sqrt{0.196^2 + 0.149^2} = 1.54$. Note that since the CCE$_{\mathrm{A}}$ and CCE estimates are based on different samples, it is natural to assume that their covariance equals zero.

[22] A similar outcome is also reported by Cui et al. (2020)

CCE approach by Pesaran (2006). However, to date this rank condition could not be verified as it relates to the rank of the unobserved matrix of factor loadings. In practice, the rank condition is therefore typically assumed to hold.

In this paper we have outlined a straightforward procedure to evaluate whether the rank condition holds in the model of interest given a chosen set of cross-sectional averages. If the rank condition is found not to hold, the procedure can be applied in an augmentation strategy, combined with an Information Criterion, to determine the set of CSA that restores the rank condition. In practice, our approach is therefore generally applicable to check whether the chosen cross-section averages are sufficient to satisfy the rank condition or whether additional variables should be explored. This property was confirmed in simulation experiments and illustrated by analyzing the impact of the Dodd-Frank Act on the profitability of U.S. banking institutions.

References

Abadir, K. M. and Magnus, J. R. (2005). Matrix Algebra. Cambridge University Press, Cambridge.

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

Al-Sadoon, M. M. (2017). A unifying theory of tests of rank. Journal of Econometrics, 199(1):49–62.

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. Statistics & Probability Letters, 80(23):1806 – 1813.

Athanasoglou, P. P., Brissimis, S. N., and Delis, M. D. (2008). Bank-specific, industry-specific and macroeconomic determinants of bank profitability. Journal of International Financial Markets, Institutions and Money, 18:121–136.

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

Baily, M. N., Klein, A., and Schardin, J. (2020). The Impact of the Dodd- Frank Act on Financial Stability and Economic Growth. rsf: the russell sage foundation journal of the social sciences, 3:20–47.

Baker, M. and Wurgler, J. (2015). Do Strict Capital Requirements Raise the Cost of Capital? Bank Regulation, Capital Structure, and the Low-Risk Anomaly. American Economic Review, 105(5):315–320.

Bordo, M. D. and Duca, J. V. (2018). The impact of the dodd-frank act on small business. Staff Reports 1806, Federal Reserve Bank of New York.

Camba-Mendez, G. and Kapetanios, G. (2009). Statistical tests and estimators of the rank of a matrix and their applications in econometric modelling. Econometric Reviews, 28(6):581–611.

Cetorelli, N. and Traina, J. (2018). Resolving "too big to fail". Staff Reports 859, Federal Reserve Bank of New York.

Chudik, A. and Pesaran, M. H. (2015). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. Journal of Econometrics, 188(2):393 – 420. Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao.

Cragg, J. G. and Donald, S. G. (1997). Inferring the rank of a matrix. Journal of Econometrics, 76(1-2):223–250.

Cui, G., Sarafidis, V., and Yamagata, T. (2020). Iv estimation of spatial dynamic panels with interactive effects: Large sample theory and an application on bank attitude toward risk. Working Paper Series 11/20, Department of Econometrics and Business Statistics at Monash University.

De Vos, I. and Everaert, G. (2021). Bias-corrected common correlated effects pooled estimation in dynamic panels. Journal of Business & Economic Statistics, 39(1):294–306.

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

Fan, J. and Liao, Y. (2020). Learning latent factors from diversified projections and its applications to over-estimated and weak factors. Journal of the American Statistical Association.

Fernandez, V. (2011). Spatial linkages in international financial markets. Quantitative Finance, 11:237–245.

Gao, Y., Liao, S., and Wang, X. (2018). Capital markets' assessment of the economic impact of the dodd-frank act on systemically important financial firms. Journal of Banking & Finance, 86:204–223.

Goddard, J., Liu, H., Molyneux, P., and Wilson, J. O. S. (2011). The persistence of bank profit. Journal of Banking & Finance, 35:2881–2890.

Goddard, J., Molyneux, P., and Wilson, J. O. S. (2004). The profitability of european banks: a crosssectional and dynamic panel analysis. The Manchester School, 72(3):363–381.

Gropp, R. and Vesala, J. (2004). Deposit insurance, moral hazard and market monitoring. Review of Finance, 8:571–602.

Hakenes, H. and Schnabel, I. (2011). Bank size and risk taking under Basel II. Journal of Banking and Finance, 35:1436–1449.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels. Journal of Econometrics, 93:345–368.

Harding, M. and Lamarche, C. (2011). Least squares estimation of a panel data model with multifactor error structure and endogenous covariates. Economics Letters, 111(3):197–199.

Harding, M., Lamarche, C., and Pesaran, M. (2018). Common correlated effects estimation of heterogeneous: Dynamic panel quantile regression models. USC Dornsife INET - Research Papers Series 18-11, Bank of Canada.

Iannotta, G., Nocera, G., and Sironi, A. (2007). Ownership structure, risk and performance in the European banking industry. Journal of Banking & Finance, 31:2127–2149.

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. Journal of Business & Economic Statistics, forthcoming.

Juodis, A. and Sarafidis, V. (2021). An incidental parameters free inference approach for panels with common shocks. Journal of Econometrics, forthcoming.

Kapetanios, G. (2008). A bootstrap procedure for panel data sets with many cross-sectional units. Econometrics Journal, 11(2):377–395.

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. Journal of Business and Economic Statistics, 28(3):397–409.

Kapetanios, G., Pesaran, M., and Yamagata, T. (2011). Panels with non-stationary multifactor error structures. Journal of Econometrics, 160(2):326–348.

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.

Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. Journal of Econometrics, 133(1):97–126.

Lee, C. and Hsieh, M. (2013). The impact of bank capital on profitability and risk in Asian banking. Journal of International Money and Finance, 32:251–281.

Mattana, P., Petroni, F., and Rossi, S. P. S. (2015). A test for the too-big-to-fail hypothesis for European banks during the financial crisis. Applied Economics, 47:319–332.

Morgan, D. P. and Stiroh, K. J. (2005). Too big to fail after all these years. Staff Reports 220, Federal Reserve Bank of New York.

Norkute, M., Sarafidis, V., Yamagata, T., and Cui, G. (2020). Instrumental variable estimation of dynamic linear panel data models with defactored regressors and a multifactor error structure. Journal of Econometrics, forthcoming.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. The Review of Economics and Statistics, 92(4):1004–1016.

Pesaran, M. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. Econometrica, 74(4):967–1012.

Pesaran, M. H., Smith, L. V., and Yamagata, T. (2007). Panel unit root tests in the presence of a multifactor error structure. IZA Discussion Papers 3254, Institute for the Study of Labor (IZA).

Robin, J.-M. and Smith, R. J. (2000). Tests of rank. Econometric Theory, 16(2):151–175.

Sarafidis, V. and Wansbeek, T. (2012). Cross-Sectional Dependence in Panel Data Analysis. Econometric Reviews, 31(5):483–531.

Sironi, A. (2003). Testing for market discipline in the European banking industry: evidence from subordinated debt issues. Journal of Money, Credit and Banking, 35:443–472.

Staikouras, C. K. and Wood, G. E. (2004). The Determinants Of European Bank Profitability. International Business & Economics Research Journal, 3(6):57–68.

Stern, G. H. and Feldman, R. J. (2009). Too big to fail: The hazards of bank bailouts. Washington, DC. ISBN: 0-8157-8152-0 220, Brookings Institution Press.

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

Trapani, L. (2018). A randomized sequential procedure to determine the number of factors. Journal of the American Statistical Association, 113(523):1341–1349.

Völz, M. and Wedow, M. (2011). Market discipline and too-big-to-fail in the CDS market: does banks size reduce market discipline? Journal of Empirical Finance, 18:195–210.

Wang, Q. (2015). Fixed-effect panel threshold model using Stata. The Stata Journal, 15:121–134.

Westerlund, J. and Urbain, J. (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.

Zhu, H., Sarafidis, V., and Silvapulle, M. (2020). A new structural break test for panels with common factors. The Econometrics Journal, 23:137–155.

Zimmerman, G. (1996). Factors Influencing Community Bank Performance in California. FRBSF ECONOMIC REVIEW, 1:26–42.

Appendices

Appendix A Proofs of theoretical results

A.1 Proof of Theorem 1

Let $\mathbf{M}$ be a given $T \times n$ matrix $(T > n)$ and let $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n]'$ be an $n \times T$ random matrix, where $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ are i.i.d.$MN(\mathbf{0}, \mathbf{I}_T)$ in $\mathbb{R}^T$.

Part (i). We wish to prove that

$$\Pr[\operatorname{rank}(\boldsymbol{\Psi}\mathbf{M}) = \operatorname{rank}(\mathbf{M})] = 1.$$

Case 1. $\mathbf{M}$ has full column rank, i.e., $\operatorname{rank}(\mathbf{M}) = n$.

Consider the row-matrix representation of a product of two matrices:

$$\boldsymbol{\Psi}\mathbf{M} = \begin{bmatrix} \boldsymbol{\psi}_1'\mathbf{M} \\ \boldsymbol{\psi}_2'\mathbf{M} \\ \vdots \\ \boldsymbol{\psi}_n'\mathbf{M} \end{bmatrix}.$$

It suffices to show that

$$\begin{aligned} \Pr\left[\left\{\boldsymbol{\psi}_1'\mathbf{M}, \ldots, \boldsymbol{\psi}_n'\mathbf{M}\right\} \text{ are linearly independent}\right] &= 1 \Leftrightarrow \\ \Pr\left[\left\{\mathbf{M}'\boldsymbol{\psi}_1, \ldots, \mathbf{M}'\boldsymbol{\psi}_n\right\} \text{ are linearly independent}\right] &= 1. \end{aligned} \tag{A-1}$$

Let $\mathbf{z}_i = \mathbf{M}'\boldsymbol{\psi}_i$ denote an $n \times 1$ vector, for $i = 1, \ldots, n$. Since $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ are i.i.d.$MN(\mathbf{0}, \mathbf{I}_T)$, it follows that $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are i.i.d.$MN(\mathbf{0}, \mathbf{M}'\mathbf{M})$, with $\mathbf{M}'\mathbf{M}$ non-singular because $\mathbf{M}$ has full rank. Let $\widetilde{\mathbf{z}}_i$ denote a specific realization of $\mathbf{z}_i$, where $\widetilde{\mathbf{z}}_i \in \mathbb{R}^n$. Define $\mathbf{y} = \operatorname{vec}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and $\widetilde{\mathbf{y}} = \operatorname{vec}(\widetilde{\mathbf{z}}_1, \ldots, \widetilde{\mathbf{z}}_n)$, both $n^2 \times 1$ vectors. Let

$$A = \left\{\widetilde{\mathbf{y}} \in \mathbb{R}^{n^2} : \widetilde{\mathbf{z}}_1, \ldots, \widetilde{\mathbf{z}}_n \text{ are linearly dependent}\right\}.$$

We have

$$\begin{aligned} & \Pr[\mathbf{z}_1, \ldots, \mathbf{z}_n \text{ are linearly dependent}] \\ = \; & E(I[\mathbf{y} \in A]) \\ = \; & E(E(I[\mathbf{y} \in A] \,|\mathbf{z}_1, \ldots, \mathbf{z}_{n-1})) = 0 \end{aligned} \tag{A-2}$$

because

$$E(I[\mathbf{y} \in A] \,|\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}) = 0.$$

Therefore, we have proved that

$$\Pr\left[\mathbf{z}_1,\ldots,\mathbf{z}_n \text{ are linearly dependent}\right]=0,$$

and thereby

$$\Pr\left[\operatorname{rank}\left(\boldsymbol{\Psi}\mathbf{M}\right)=\operatorname{rank}\left(\mathbf{M}\right)=n\right]=1.$$

Case 2. $\mathbf{M}$ has less than full column rank, i.e., $\operatorname{rank}(\mathbf{M})=n_1<n$.

Partition $\mathbf{M}=[\mathbf{M}_1 \vdots \mathbf{M}_2]$, where $\mathbf{M}_1$ is $T\times n_1$ and $\mathbf{M}_2$ is $T\times(n-n_1)$, such that $\operatorname{rank}(\mathbf{M}_1)=n_1$. Similarly, partition $\boldsymbol{\Psi}$ such that

$$\boldsymbol{\Psi}=\left[\begin{array}{c}\boldsymbol{\Psi}_1\\ \hdashline \boldsymbol{\Psi}_2\end{array}\right]$$

where $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are $n_1\times T$ and $(n-n_1)\times T$ respectively with $\operatorname{rank}(\boldsymbol{\Psi}_1)=n_1$ with probability 1. Thus, the product between $\boldsymbol{\Psi}$ and $\mathbf{M}$ can be written as

$$\boldsymbol{\Psi}\mathbf{M}=\left[\begin{array}{c:c}\boldsymbol{\Psi}_1\mathbf{M}_1 & \boldsymbol{\Psi}_1\mathbf{M}_2\\ \hdashline \boldsymbol{\Psi}_2\mathbf{M}_1 & \boldsymbol{\Psi}_2\mathbf{M}_2\end{array}\right]_{n\times n}. \tag{A-3}$$

Based on exactly the same arguments as in **Case 1**, it can be shown that $\operatorname{rank}(\boldsymbol{\Psi}_1\mathbf{M}_1)=n_1$ with probability 1. However, since $\boldsymbol{\Psi}_1\mathbf{M}_1$ is a submatrix of $\boldsymbol{\Psi}\mathbf{M}$, $\operatorname{rank}(\boldsymbol{\Psi}\mathbf{M})\geq n_1$. Therefore, we have

$$n_1=\operatorname{rank}\left(\boldsymbol{\Psi}_1\mathbf{M}_1\right)\leq\operatorname{rank}\left(\boldsymbol{\Psi}\mathbf{M}\right)\leq\min\left\{\operatorname{rank}\left(\boldsymbol{\Psi}\right),\operatorname{rank}\left(\mathbf{M}\right)\right\}=n_1.$$

Hence, $\operatorname{rank}(\boldsymbol{\Psi}\mathbf{M})=n_1=\operatorname{rank}(\mathbf{M})$ with probability 1. This completes part (i) of the theorem.

Part (ii). Note that we can write by simple addition and subtraction

$$T^{-1/2}\boldsymbol{\Psi}\overline{\mathbf{Z}}=T^{-1/2}(\boldsymbol{\Psi}\mathbf{F}\overline{\mathbf{C}}+\boldsymbol{\Psi}\overline{\mathbf{U}})=T^{-1/2}\boldsymbol{\Psi}\mathbf{F}\mathbf{C}+T^{-1/2}\boldsymbol{\Psi}\mathbf{F}(\overline{\mathbf{C}}-\mathbf{C})+T^{-1/2}\boldsymbol{\Psi}\overline{\mathbf{U}}.$$

Recall that $\boldsymbol{\Psi}=[\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_n]'$, with its rows given by $\boldsymbol{\psi}_k\sim iidMN\left(\mathbf{0}_{T\times 1},\mathbf{I}_T\right)$ for $k=1,\ldots,n$. Consider then the kth row of $T^{-1/2}\boldsymbol{\Psi}\overline{\mathbf{U}}$, and write it as $T^{-1/2}\sum_{t=1}^{T}\psi_{kt}\overline{\mathbf{u}}_t'$, with $\boldsymbol{\psi}_k=[\psi_{k1},\ldots,\psi_{kT}]'$ and $\overline{\mathbf{U}}=[\overline{\mathbf{u}}_1,\ldots,\overline{\mathbf{u}}_T]'$. By the independence of $\boldsymbol{\Psi}$ and $\overline{\mathbf{U}}$ we have for every $k=1,\ldots,n$ that $E(\sum_{t=1}^{T}\psi_{kt}\overline{\mathbf{u}}_t')=\mathbf{0}_{1\times n}$ and

$$\operatorname{Var}\left(\frac{\sum_{t=1}^{T}\psi_{kt}\overline{\mathbf{u}}_t'}{\sqrt{T}}\right)=\frac{1}{T}\sum_{t=1}^{T}\sum_{t'=1}^{T}E\left(\psi_{kt}\psi_{kt'}\right)E\left(\overline{\mathbf{u}}_t\overline{\mathbf{u}}_{t'}'\right)=\frac{1}{T}\sum_{t=1}^{T}E\left(\overline{\mathbf{u}}_t\overline{\mathbf{u}}_t'\right)=O(N^{-1}),$$

because $E\left(\psi_{kt}\psi_{kt'}\right)=0$ for $t'\neq t$, $E\left(\psi_{kt}\psi_{kt}\right)=E\left(\psi_{kt}^2\right)=1$ and $E(\|\overline{\mathbf{u}}_t\|^2)=O\left(N^{-1}\right)$ by A.4 of Lemma 1 in Pesaran (2006) under Assumptions 1 and 5. Hence, $\left\|T^{-1/2}\boldsymbol{\Psi}\overline{\mathbf{U}}\right\|=O_p(N^{-1/2})$ as $(N,T)\to\infty$.

Consider next $T^{-1/2}\mathbf{\Psi F}$. By the independence of $\mathbf{\Psi}$ and $\mathbf{F}$ we have $E\left(\mathbf{\Psi F}\right) = \mathbf{0}_{n\times m}$, and since $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_T]'$ we can write the k-th row of $\mathbf{\Psi F}$ as $\sum_{t=1}^{T} \psi_{kt}\mathbf{f}_t'$ such that

$$\text{Var}\left(\frac{\sum_{t=1}^{T} \psi_{kt}\mathbf{f}_t'}{\sqrt{T}}\right) = \frac{1}{T}\sum_{t=1}^{T} E\left(\psi_{kt}\psi_{kt}\right) E\left(\mathbf{f}_t\mathbf{f}_t'\right) = \frac{1}{T}\sum_{t=1}^{T} E\left(\mathbf{f}_t\mathbf{f}_t'\right) = O(1),$$

because $E\left(\mathbf{f}_t\mathbf{f}_t'\right) = O(1)$ for every t (Assumption 2). Hence, we have $\left\|T^{-1/2}\mathbf{\Psi F}\right\| = O_p\left(1\right)$. Noting then that $\left\|\overline{\mathbf{C}} - \mathbf{C}\right\| = O_p(N^{-1/2})$ under Assumption 3, it follows that

$$\left\|T^{-1/2}\mathbf{\Psi F}(\overline{\mathbf{C}} - \mathbf{C})\right\| \leq \left\|T^{-1/2}\mathbf{\Psi F}\right\| \left\|\overline{\mathbf{C}} - \mathbf{C}\right\| = O_p(N^{-1/2})$$

Hence, combining the results above, as $(N, T) \to \infty$

$$T^{-1/2}\mathbf{\Psi}\overline{\mathbf{Z}} = T^{-1/2}\mathbf{\Psi FC} + T^{-1/2}\mathbf{\Psi F}(\overline{\mathbf{C}} - \mathbf{C}) + T^{-1/2}\mathbf{\Psi}\overline{\mathbf{U}} = T^{-1/2}\mathbf{\Psi FC} + O_p(N^{-1/2})$$

where also $\left\|T^{-1/2}\mathbf{\Psi FC}\right\| \leq \left\|T^{-1/2}\mathbf{\Psi F}\right\| \left\|\mathbf{C}\right\| = O_p(1)$ since $\left\|\mathbf{C}\right\| < \infty$ under Assumption 3. Hence, the proof of part (ii) of the theorem is complete. □

Appendix B Additional simulation results

Table 8: Algorithm 1: Selection percentages for expansion CSA

		$\overline{\mathbf{Z}}_{w,1}$				$\overline{\mathbf{Z}}_{w,2}$				$\overline{\mathbf{Z}}^{(e)}$				$\overline{\mathbf{Z}}^{(g)}$			
	(N,T)	20	50	100	200	20	50	100	200	20	50	100	200	20	50	100	200
Experiment 1	20	2	0	0	0	1	0	0	0	2	0	0	0	1	0	0	0
	50	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
	100	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
	200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Experiment 2	20	37	42	46	51	33	43	48	47	23	10	1	0	0	0	0	0
	50	34	39	43	47	28	40	50	52	32	19	4	0	0	0	0	0
	100	32	32	41	50	27	30	47	49	35	38	11	1	1	0	0	0
	200	27	32	35	46	26	30	33	47	42	37	32	7	1	0	0	0
	500	22	25	34	39	27	29	32	39	42	46	34	22	0	0	0	0
	1000	21	25	29	27	22	26	25	35	50	49	46	38	1	0	0	0
Experiment 3	20	4	1	0	0	3	0	0	0	93	98	100	100	1	0	0	0
	50	3	0	0	0	2	0	0	0	96	99	100	100	2	0	0	0
	100	1	0	0	0	2	0	0	0	96	100	100	100	1	0	0	0
	200	1	0	0	0	1	0	0	0	96	100	100	100	2	0	0	0
	500	0	0	0	0	0	0	0	0	98	99	100	100	1	0	0	0
	1000	1	0	0	0	1	0	0	0	98	100	100	100	1	0	0	0
Experiment 3	20	26	20	19	27	23	18	24	18	0	0	0	0	2	0	0	0
$\overline{\mathbf{Z}}_{+,sub} = \overline{\mathbf{Z}}_{+} \backslash \overline{\mathbf{Z}}^{(e)}$	50	22	13	20	21	17	17	13	18	0	0	0	0	2	0	0	0
	100	22	11	12	15	16	16	15	19	0	0	0	0	2	0	0	0
	200	24	13	15	16	18	14	12	19	0	0	0	0	4	0	0	0
	500	22	11	12	16	18	11	14	19	0	0	0	0	3	1	0	0
	1000	19	14	14	18	19	14	17	16	0	0	0	0	2	0	0	0

Notes: Reported are percentages out of 2000 Monte Carlo iterations that the CSA stated in the column has been selected as an expansion by the IC given in Eq. (25). Since multiple expansions can be selected on each sample size, the percentages do not necessarily sum to 100. The bottom panel displays selection frequencies in Experiment 3 when $\overline{\mathbf{Z}}^{(e)}$ is not a selectable option. That is, the set of proposal expansions is $\overline{\mathbf{Z}}_{+,sub} = \{\overline{\mathbf{Z}}_{w,1}, \overline{\mathbf{Z}}_{w,2}, \overline{\mathbf{Z}}^{(g)}\}$.

Table 9: Algorithm 1: Sensitivity and Specificity

		RC satisfied rate				Sensitivity				Specificity			
	(N, T)	20	50	100	200	20	50	100	200	20	50	100	200
Experiment 1	20	1.00	1.00	1.00	1.00	0.65	0.74	0.67	0.67	1.00	1.00	1.00	1.00
	50	1.00	1.00	1.00	1.00	0.73	0.81	0.80	0.81	1.00	1.00	1.00	1.00
	100	1.00	1.00	1.00	1.00	0.83	0.80	0.84	0.86	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	0.85	0.88	0.89	0.91	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	0.87	0.92	0.92	0.92	1.00	1.00	1.00	1.00
	1000	1.00	1.00	1.00	1.00	0.89	0.93	0.93	0.95	1.00	1.00	1.00	1.00
Experiment 2	20	0.92	0.95	0.96	0.97	0.80	0.94	0.98	0.98	0.00	0.00	0.00	0.00
	50	0.94	0.98	0.96	0.99	0.92	0.96	0.99	1.00	0.00	0.00	0.00	0.00
	100	0.93	0.99	1.00	1.00	0.96	1.00	1.00	1.00	0.00	0.00	1.00	1.00
	200	0.94	0.99	1.00	1.00	0.96	1.00	1.00	1.00	0.00	0.00	1.00	0.00
	500	0.91	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.00	0.00	1.00	1.00
	1000	0.92	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.00	1.00	1.00	1.00
Experiment 3	20	0.92	0.97	0.99	0.99	0.36	0.39	0.38	0.40	0.63	0.95	0.22	0.00
	50	0.96	0.99	0.99	1.00	0.64	0.68	0.67	0.73	0.94	0.52	0.00	1.00
	100	0.96	0.99	1.00	1.00	0.85	0.93	0.96	0.91	1.00	1.00	1.00	1.00
	200	0.96	1.00	1.00	1.00	0.91	0.98	0.98	0.96	0.89	1.00	1.00	1.00
	500	0.98	0.99	1.00	1.00	0.93	0.99	0.99	0.97	1.00	1.00	1.00	1.00
	1000	0.98	1.00	1.00	1.00	0.93	0.99	0.99	0.99	1.00	1.00	1.00	1.00
Experiment 3	20	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.92	0.96	0.92	0.94
$\overline{\mathbf{Z}}_{+,sub} = \overline{\mathbf{Z}}_{+} \backslash \overline{\mathbf{Z}}^{(e)}$	50	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.97	0.97	0.98	0.99
	100	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.98	0.99	0.99	1.00
	200	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.97	0.99	0.99	0.98
	500	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.96	0.99	1.00	1.00
	1000	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99	0.99

Notes: (i) Reported in the left panel is the fraction of MC samples where the rank condition $(\varrho = m)$ is satisfied (restored) after application of Algorithm 1. The middle panel displays the 'Sensitivity' of the RC classifier, or the rate of correctly obtaining $\widehat{RC} = 1$ for the cases where the RC is satisfied/restored $(\frac{\text{\#true RC=1 conclusions}}{\text{\#true RC=1 conclusions+\#false RC=0 conclusions}})$, and the right-most panel gives the 'Specificity', or the rate of correctly obtaining $\widehat{RC} = 0$ when the RC is indeed violated/not restored $(\frac{\text{\#true RC=0 conclusions}}{\text{\#true RC=0 conclusions+\#false RC=1 conclusions}})$. Note that when there are no $RC = 0$ cases and also no $\widehat{RC} = 0$ conclusions, then Specificity = 1, and similarly for the Sensitivity. The inverse of Sensitivity and Specificity give respectively the false positive (false RC holds conclusions) and false negative rates (false RC violated conclusions). (ii) The RC classifier employs the GR estimator with $m_{max} = 7$ to estimate m, and the rank estimator employs the random projection with $\alpha_N = 20\alpha N^{-1}$. (iii) The bottom panel gives outcomes for Algorithm 1 when the rank-restoring expansion CSA $\overline{\mathbf{Z}}^{(e)}$ is not among the set of proposal expansions such that it is impossible to restore the RC. (iv) Note that the classifier Sensitivity/Specificity are not separately reported when evaluating the RC based on $\overline{\mathbf{Z}}$, because they are identical to the 'Classification Accuracy' reported in the main text. That is, when the RC is satisfied for $\overline{\mathbf{Z}}$ (experiment 1), then Specificity=1 and Sensitivity equals the classification accuracy reported in table 2. Conversely, when RC is violated for $\overline{\mathbf{Z}}$, then Sensitivity=1 and Specificity is the reported accuracy.

B.1 Estimation results for $\beta = 3$

Table 10: Estimation results for $\beta = 3$ in Experiment 1

		bias				*rmse*			
	(N, T)	20	50	100	200	20	50	100	200
CCE	20	0.053	0.051	0.051	0.047	0.120	0.089	0.075	0.059
	50	0.021	0.020	0.021	0.023	0.065	0.048	0.040	0.032
	100	0.011	0.011	0.011	0.013	0.050	0.032	0.024	0.020
	200	0.005	0.004	0.007	0.006	0.034	0.021	0.017	0.013
	500	0.002	0.001	0.002	0.003	0.020	0.013	0.010	0.007
	1000	0.002	0.002	0.001	0.001	0.015	0.009	0.007	0.005
CCE_A	20	0.052	0.051	0.051	0.047	0.118	0.089	0.075	0.059
	50	0.021	0.020	0.021	0.023	0.065	0.048	0.040	0.032
	100	0.011	0.011	0.011	0.013	0.051	0.032	0.024	0.020
	200	0.005	0.004	0.007	0.006	0.034	0.021	0.017	0.013
	500	0.002	0.001	0.002	0.003	0.020	0.013	0.010	0.007
	1000	0.002	0.002	0.001	0.001	0.015	0.009	0.007	0.005

Note: Reported are estimation bias for β and root mean square error (rmse).

Table 11: Estimation results for $\beta = 3$ in Experiment 2

		bias				*rmse*			
	(N, T)	20	50	100	200	20	50	100	200
CCE	20	0.819	0.839	0.839	0.844	0.826	0.842	0.841	0.846
	50	0.815	0.839	0.849	0.851	0.820	0.841	0.850	0.852
	100	0.825	0.835	0.848	0.853	0.830	0.837	0.849	0.854
	200	0.825	0.845	0.850	0.854	0.829	0.847	0.851	0.854
	500	0.824	0.832	0.849	0.852	0.828	0.834	0.850	0.852
	1000	0.823	0.841	0.848	0.852	0.826	0.843	0.848	0.852
CCE_A	20	0.090	0.068	0.062	0.043	0.246	0.201	0.186	0.138
	50	0.063	0.029	0.040	0.022	0.219	0.137	0.162	0.101
	100	0.066	0.016	0.006	0.006	0.237	0.080	0.022	0.016
	200	0.057	0.016	0.005	0.006	0.211	0.104	0.017	0.058
	500	0.082	0.003	0.002	0.002	0.263	0.049	0.010	0.006
	1000	0.069	0.001	0.001	0.001	0.237	0.009	0.007	0.005

Note: Reported are estimation bias for β and root mean square error (rmse).

Table 12: Estimation results for $\beta = 3$ in Experiment 3

		bias				*rmse*			
	(N, T)	20	50	100	200	20	50	100	200
CCE	20	0.655	0.677	0.680	0.682	0.672	0.685	0.687	0.687
	50	0.670	0.689	0.694	0.698	0.679	0.695	0.698	0.701
	100	0.669	0.689	0.698	0.707	0.680	0.695	0.701	0.708
	200	0.678	0.694	0.704	0.705	0.687	0.700	0.707	0.707
	500	0.681	0.682	0.701	0.711	0.691	0.687	0.705	0.713
	1000	0.671	0.690	0.700	0.705	0.680	0.694	0.702	0.707
CCE$_{\text{A}}$	20	0.051	0.033	0.029	0.020	0.149	0.083	0.060	0.041
	50	0.026	0.010	0.012	0.010	0.096	0.059	0.050	0.024
	100	0.015	0.009	0.006	0.007	0.069	0.040	0.022	0.017
	200	0.016	0.002	0.004	0.003	0.079	0.021	0.016	0.012
	500	0.006	0.002	0.001	0.002	0.056	0.035	0.010	0.006
	1000	0.008	0.001	0.000	0.000	0.055	0.009	0.007	0.005
CCE$_{\text{A},sub}$	20	0.504	0.544	0.535	0.535	0.539	0.574	0.566	0.566
	50	0.556	0.591	0.590	0.581	0.580	0.614	0.613	0.604
	100	0.550	0.592	0.604	0.597	0.577	0.616	0.627	0.622
	200	0.548	0.602	0.614	0.600	0.575	0.625	0.635	0.622
	500	0.551	0.604	0.614	0.600	0.578	0.625	0.635	0.625
	1000	0.549	0.592	0.594	0.591	0.577	0.616	0.619	0.617

Note: Reported are estimation bias for β and root mean square error (rmse). CCE$_{\text{A},sub}$ denotes the outcome of Algorithm 1 with the set of potential augmentations given by $\overline{\mathbf{Z}}_{+,sub} = \{\overline{\mathbf{Z}}_{w,1}, \overline{\mathbf{Z}}_{w,2}, \overline{\mathbf{Z}}^{(g)}\}$ in stead of $\overline{\mathbf{Z}}_{+}$.