Journal of Econometrics 235 (2023) 1981–2004

Contents lists available at ScienceDirect

Journal of Econometrics

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

One-way or two-way factor model for matrix sequences?

Yong He [a], Xinbing Kong [b,*], Lorenzo Trapani [c], Long Yu [d]

[a] Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan, 250100, China
[b] Nanjing Audit University, Nanjing, 211815, China
[c] University of Leicester, United Kingdom of Great Britain, United Kingdom
[d] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China

ARTICLE INFO

Article history:
Received 23 December 2021
Received in revised form 11 February 2023
Accepted 23 February 2023
Available online 14 March 2023

Keywords:
Matrix sequence
Matrix factor model
Principal component analysis
Projection estimation
Randomised tests

ABSTRACT

This paper investigates the issue of determining the dimensions of row and column factor spaces in matrix-valued data. Exploiting the eigen-gap in the spectrum of sample second moment matrices of the data, we propose a family of randomised tests to check whether a one-way or two-way factor structure exists or not. Our tests do not require any arbitrary thresholding on the eigenvalues, and can be applied with (virtually) no restrictions on the relative rate of divergence of the cross-sections to the sample sizes as they pass to infinity. Although tests are based on a randomisation which does not vanish asymptotically, we propose a de-randomised, "strong" (based on the Law of the Iterated Logarithm) decision rule to choose in favour or against the presence of common factors. We use the proposed tests and decision rule in two ways. We further cast our individual tests in a sequential procedure whose output is an estimate of the number of common factors. Our tests are built on two variants of the sample second moment matrix of the data: one based on a row (or column) "flattened" version of the matrix-valued sequence, and one based on a projection-based method. Our simulations show that both procedures work well in large samples and, in small samples, the one based on the projection method delivers a superior performance compared to existing methods in virtually all cases considered.

1. Introduction

Matrix time series can be defined as a sequence of $p_1 \times p_2$ random matrices $\{X_t, 1 \leq t \leq T\}$, with each random matrix used to model observations that are well structured to be an array. Such datasets are of great interest in a wide variety of applied sciences in general, and in social sciences in particular. For example, in macroeconomics a "classical" application of matrix-valued time series (see the recent paper by Chen et al., 2022, and the discussion therein) is modelling the import-export volumes between countries for one product family such as e.g. chemical, food, or machinery and electronic. In this example, also known as a "dynamic transport network", at each point in time one can construct a matrix where the columns represent imports into a country and the rows exports towards a country (with the main diagonal of course empty). Another possible example, studied in Wang et al. (2019), is a matrix of time series whose rows contain some macroeconomic indicator (GDP, inflation, interest rates...) and whose columns represent different countries. Further, in the context of financial data, Wang et al. (2019) study a matrix-valued time-series of portfolio returns where each portfolio

* Corresponding author.
E-mail addresses: heyong@sdu.edu.cn (Y. He), xinbingkong@126.com (X. Kong), Lorenzo.Trapani@nottingham.ac.uk (L. Trapani), fduyulong@163.com (L. Yu).

https://doi.org/10.1016/j.jeconom.2023.02.008
0304-4076/

is identified by a size level and by a BE ratio level; in the same paper, another example is provided, conceptually similar to the one based on macroeconomic indicators described above, where a matrix-valued time series is considered with different companies on each row, and different company financials on each column. Finally, in marketing studies, a very promising application is to time series of customers' ratings on a large number of items in an online platform; a well-known application of such a "recommender system" (Koren et al., 2009) is where, as time elapses, several customers are asked to express their level of satisfaction with several movies/TV shows. We also refer to the papers by Chen and Fan (2021) and Gao et al. (2021) for further discussion and examples ranging from health sciences (such as electronic health records and ICU data), to 2-D image data processing.

When dealing with such complex datasets, exploring the possibility of dimensionality reduction is of pivotal importance. A possible way of achieving this is to "flatten" the data, and model the vectorised sequence of matrices as

$$\underset{p_1p_2\times 1}{\text{Vec}\,(X_t)} = \underset{p_1p_2\times k}{\Lambda}\ \underset{k\times 1}{f_t} + \underset{p_1p_2\times 1}{u_t}\ , \tag{1.1}$$

where f_t is a (low-dimensional) vector of common factors. Such a modelling strategy has been studied, in the context of vector-valued series, in numerous contributions, and we refer to Bai and Wang (2016) for a comprehensive review. Although (1.1) does lead to dimension reduction, further refinements may still be desirable. On the one hand, (1.1) requires the estimation of p_1p_2k parameters in the loading matrix Λ; this number may still be too large in empirical applications, in particular when the cross-sectional dimensions p_1 and p_2 are large. On the other hand, given that X_t is a sequence of matrices, a better modelling approach could be based on allowing for the presence of common factors along the rows and along the columns of X_t, rather than destroying the matrix nature of the data by vectorising it.

1.1. The two-way factor model for matrix-valued time series

In order to make full use of the matrix structure, a parsimonious modelling approach has been proposed in a recent, seminal paper by Wang et al. (2019), who assume that X_t is driven by a low-dimensional set of common factors across the row and column dimensions:

$$\underset{p_1\times p_2}{X_t} = \underset{p_1\times k_1}{R}\ \underset{k_1\times k_2}{F_t}\ \underset{k_2\times p_2}{C'} + \underset{p_1\times p_2}{E_t}\ ,\ 1\le t\le T,\ k_1,\ k_2>0. \tag{1.2}$$

In (1.2), R is the $p_1\times k_1$ row factor loading matrix explaining the variation of X_t across the rows, C is the $p_2\times k_2$ column factor loading matrix reflecting the differences across the columns of X_t, F_t is the common factor matrix, and E_t is an idiosyncratic component. At a glance, a natural competitor of (1.2) could be a group (vector) factor model (see e.g. Ando and Bai, 2016, and Andreou et al., 2019): in such a class of models, there is only one cross-section, and this one cross-section contains variables of the same nature (say, considering an example above, the set of macroeconomic indicators) which are well-grouped with known or unknown group membership. The common factors are organised into groups, and the interrelations within and between groups are characterised by such factors. Conversely, the data X_t in (1.2) are genuinely matrix-valued, with two cross-sectional dimensions of different nature (considering the examples mentioned above, these could be countries and macroeconomic indicators in the context of macroeconomic data; or customers and commodities in recommending systems). Hence, the common components in the matrix factor models reflect the interplay between the two different cross-sections: for example, in the context of recommending systems, ratings are high whenever the purchasers' consumption preferences (rows in R) match the underlying characteristics of items displayed online (rows in C), thus (1.2) is a natural modelling of the interactive effect between the row and column cross sections. In this context, it is natural to expect that (1.2), which takes the matrix nature of the data into account, is a better approach than using models based on vectorising X_t, where the presence of groups arises from artificially stacking the columns (rows) of matrix-valued data. Moreover, (1.2) has the added bonus of reducing the dimensionality compared to a model like (1.1): whilst in the latter case one needs to estimate p_1p_2k coefficients, in the case of (1.2) such parameter complexity is reduced to $p_1k_1+p_2k_2$. This can be viewed even more neatly if one considers the following alternative version of (1.2)

$$\text{Vec}\,(X_t) = (C\otimes R)\,\text{Vec}\,(F_t) + \text{Vec}\,(E_t)\,, \tag{1.3}$$

where "$\otimes$" denotes the Kronecker product. Eq. (1.3) shows that the loadings associated with the factor structure in X_t satisfy a Kronecker product structure, whence the higher parsimony of (1.2).

In order to better understand the nature of (1.3), we consider the following example, where – as also mentioned above – X_t represents a $p_1\times p_2$ time series whose columns contain some macroeconomic indicators (GDP, inflation, interest rates...), and whose rows represent different countries.[1] Consider the following notation: $R=(\alpha_{p_1\times r}, \widetilde{R}_{p_1\times(k_1-r-l)}, \mathbf{1}_{p_1\times l})$, $C=(\mathbf{1}_{p_2\times r}, \widetilde{C}_{p_2\times(k_2-r-l)}, \gamma_{p_2\times l})$ and $F_t=\text{diag}\{(g_t)_{r\times r}, (\widetilde{F}_t)_{(k_1-r-l)\times(k_2-r-l)}, (h_t)_{l\times l}\}$. Then, (1.3) becomes

$$X_t = \alpha g_t \mathbf{1}_{r\times p_2} + \mathbf{1}_{p_1\times l} h_t \gamma' + \widetilde{R}\widetilde{F}_t\widetilde{C}' + E_t. \tag{1.4}$$

[1] In Section A.1 of the Supplementary Material, we also discuss another example, based on a similar discussion in Wang et al. (2019), which also illustrates the relationship between (1.3) and a multilevel factor model.

In (1.4), g_t and h_t are the common factors along the row and column cross-sections, respectively, and α and γ represent their loadings; $\widetilde{R}\widetilde{F}_t\widetilde{C}'$ may be viewed as an interaction effect component. Model (1.4) can be rewritten in vector form, for both the countries $j = 1, \ldots, p_1$ and the indicators $i = 1, \ldots, p_2$:

$$X_{j\cdot,t} = \alpha_{j\cdot} g_t \mathbf{1}_{r\times p_2} + \mathbf{1}_{1\times l} h_t \gamma' + \widetilde{R}_{j\cdot}\widetilde{F}_t\widetilde{C}' + E_{j\cdot,t}, \tag{1.5}$$

$$X_{\cdot i,t} = \alpha g_t \mathbf{1}_{r\times 1} + \mathbf{1}_{p_1\times l} h_t (\gamma')_{\cdot i} + \widetilde{R}\widetilde{F}_t\widetilde{C}'_{\cdot i} + E_{\cdot i,t}. \tag{1.6}$$

Eqs. (1.5) and (1.6) lend themselves to the following interpretation. The term g_t represents the common global factors affecting all countries — but the rows of α are heterogeneous, indicating that, for each country, the data have specific loadings on the global factors. Similarly, h_t are latent common factors reflecting economic states across macroeconomic indicators — but the rows of γ are heterogeneous, indicating that each macroeconomic indicator loads on the states differently. Considering the rows $X_{j\cdot,t}$ and model (1.5), the term $g_t\mathbf{1}_{r\times p_2}$ is therefore a column-invariant global factor term, while the term $\mathbf{1}_{1\times l}h_t\gamma'$ is a columnwise adjusting term. Looking at the columns $X_{\cdot i,t}$, i.e. (1.6), the term $\mathbf{1}_{p_1\times l}h_t$ contains economic state factors invariant across all indicators, while the term $\alpha g_t\mathbf{1}_{r\times 1}$ is a rowwise adjusting term. The third terms in both (1.5) and ((1.6)) reflect the interaction effect between the two cross-sections. Hence, the matrix factor model incorporates and identifies SIMULTANEOUSLY the geographical global factors common to the countries, and the economic state factors common to the indicators.

Eq. (1.4) nests several interesting special cases. Indeed, when $r = l = 1$, $\alpha = \tilde{\alpha}_{p_1\times 1}$, $\gamma = (\tilde{\gamma}_{1\times p_2})'$ and $g_t = h_t = 1$, (1.4) boils down to

$$X_t = \tilde{\alpha}_{p_1\times 1}\mathbf{1}_{1\times p_2} + \mathbf{1}_{p_1\times 1}\tilde{\gamma}_{1\times p_2} + \widetilde{R}\widetilde{F}_t\widetilde{C}' + E_t, \tag{1.7}$$

which is a model with time-invariant fixed effects along both the row and column dimensions.[2] Such fixed effects are allowed to be heterogeneous across the rows and/or columns, representing the specific effects (factors) of rows (countries) and columns (indicators); using the notation $\tilde{\alpha}_{p_1\times 1} = (r_1, \ldots, r_{p_1})'$ and $\tilde{\gamma}_{p_2\times 1} = (c_1, \ldots, c_{p_2})'$, (1.7) entails that

$$X_{ji,t} = r_j + c_i + \widetilde{R}_{j\cdot}\widetilde{F}_t\widetilde{C}'_{\cdot i} + E_{ji,t}, \tag{1.8}$$

i.e., a model with "two-way" cross-sectional fixed effects. Finally, combining (1.4) and ((1.7)) yields another special example of the matrix factor model

$$X_t = \underbrace{\alpha g_t \mathbf{1}_{r\times p_2} + \tilde{\alpha}_{p_1\times 1}\mathbf{1}_{1\times p_2}}_{I} + \underbrace{\mathbf{1}_{p_1\times l}h_t\gamma' + \mathbf{1}_{p_1\times 1}\tilde{\gamma}_{1\times p_2}}_{II} + \widetilde{R}\widetilde{F}_t\widetilde{C}' + E_t, \tag{1.9}$$

i.e. a model with: fixed effects in both the row and column dimensions, two sets of latent factors common to countries and indicators respectively, and an interaction term.

As far as inference is concerned, under (1.3) the "loadings" $C \otimes R$ can be estimated by obtaining $\widehat{C}$ and $\widehat{R}$ separately, and subsequently computing $\widehat{C} \otimes \widehat{R}$. As pointed out in Chen and Fan (2021), if one were to estimate $C \otimes R$ by ignoring the Kronecker product structure and using e.g. the standard PCA estimator studied in Bai (2003), the convergence rate of $\widehat{C \otimes R}$ in L_2-norm would be $\max\left\{T^{-1/2}, (p_1p_2)^{-1/2}\right\}$. Conversely, under (1.3), Yu et al. (2022) show that the L_2-norm convergence rates of $\widehat{C}$ and $\widehat{R}$ are, respectively

$$\max\left\{\frac{1}{\sqrt{Tp_1}}, \frac{1}{\min\{p_1, T\}\, p_2}\right\} \quad \text{and} \quad \max\left\{\frac{1}{\sqrt{Tp_2}}, \frac{1}{\min\{p_2, T\}\, p_1}\right\}.$$

Hence, $\widehat{C}\otimes\widehat{R}$ has a faster rate of convergence than $\widehat{C \otimes R}$ in the case of large dimensional datasets. Further, as far as second order properties are concerned, an estimation technique that makes full use of the dimensionality reduction implied by (1.3) is bound to result in efficiency gains. Finally, if the object of interest are C and R, direct estimation is going to be better (as well as computationally more efficient) than firstly estimating $C \otimes R$ and subsequently recovering C and R therefrom via Kronecker product decomposition (Cai et al., 2019).

As mentioned above, the first contribution to consider a factor model with a Kronecker product structure like (1.3) is Wang et al. (2019), who propose estimators of the factor loading matrices (and of the numbers of the row and column factors) based on the eigen-analysis of the auto-cross-covariance matrix. From a different perspective, and assuming cross-sectional pervasiveness along the row and column dimensions, Chen and Fan (2021) propose an estimation technique based on the eigen-analysis of a weighted average of the mean and the column (row) covariance matrix of the data; Yu et al. (2022) improve the estimation efficiency of the factor loading matrices with iterative projection algorithms. All these methodologies can also be employed to construct estimators of the number of common factors. In addition, there are also contributions which specifically address the estimation of the dimensions of the factor spaces. In the broader context of tensor factor models, Han et al. (2022) propose two approaches (one which is similar, in spirit, to the information criteria in Bai and Ng (2002), and one which is based on using the ratio of consecutive eigenvalues) to determine the dimension of the factor spaces; Lam (2021) considers estimating the number of common factors by thresholding the eigenvalues of the correlation matrix of the data (see also Chen and Lam, 2022). Further extensions and applications of the basic set-up

[2] See Kong et al. (2022), where this model is studied.

in (1.2) include the constrained version by Chen et al. (2020a) , the semiparametric estimators by Chen et al. (2020b), and the estimators developed in Chen et al. (2022); see also Han et al. (2022) . Chen and Chen (2020) apply (1.2) to the dynamic transport network in the context of international trade flows, and Chen et al. (2021b) consider applications to financial datasets.

However, even though the literature has produced several contributions to carry out inference in (1.2), no works has been done so far to seriously test the existence of the factor structure implicitly defined in (1.2). Being able to discern whether a genuine matrix factor structure exists or not is a crucial point in the analysis of matrix-valued data. As Chen and Fan (2021) put it, "[...] analyzing large scale matrix-variate data is still in its infancy, and as a result, scientists frequently analyze matrix-variate observations by separately modeling each dimension or 'flattening' them into vectors. This destroys the intrinsic multi-dimensional structure and misses important patterns in such large scale data with complex structures, and thus leads to sub-optimal results".

1.2. Hypotheses of interest and the contribution of this paper

In (1.2), both k_1 and k_2 are strictly positive, thus allowing for a collaborative dependence between the row cross-section and the column cross-section: we name this *two-way factor structure*. Since we interpret k_1 and k_2 as the numbers of row and column factors, we let $k_1 = 0$ and $k_2 = 0$ correspond to the scenarios without row factors and without column factors, respectively. When $k_2 = 0$ but $k_1 > 0$, we refer to this as having a *one-way factor structure* along the row dimension: all columns of the whole matrix sequence could be modelled by a p_1 dimensional vector factor model with effective sample size Tp_2. A similar interpretation applies to the scenario where $k_1 = 0$ and $k_2 > 0$. Finally, when $k_1 = k_2 = 0$, the matrix-valued data is simply a noise matrix.

In order to model the "boundary" cases discussed above, henceforth, we use the following CONVENTION

$$\underset{p_1\times p_2}{X_t} = \begin{cases} \underset{p_1\times k_1}{R}\ \underset{k_1\times p_2}{F_t} + \underset{p_1\times p_2}{E_t}, & k_1 > 0, k_2 = 0, \\ \underset{p_1\times k_2}{F_t}\ \underset{k_2\times p_2}{C'} + \underset{p_1\times p_2}{E_t}, & k_2 > 0, k_1 = 0, \\ \underset{p_1\times p_2}{E_t}, & k_1 = k_2 = 0, \end{cases} \tag{1.10}$$

where the first case refers to a one-way factor model along the row dimension (all columns form a vector factor model), the second case is a one-way factor model along the column dimension (all rows form a vector factor model), and the third case means absence of any factor structure. We note that – since a factor structure is well-defined only if the dimension of the factor space is finite – in (1.10) we prefer to avoid the notation $k_2 = p_2$ and $C = I_{p_2}$ (resp. $k_1 = p_1$ and $R = I_{p_1}$), even though it is mathematically equivalent to the first (resp. the second) case in (1.10).

In the context of (1.2) and (1.10), several questions naturally arise: *is there a common, latent factor structure in the rows and/or columns of X_t? How many row and/or column factors are there?* Considering the macroeconomic example discussed above, this entails checking the existence of country and/or index factors, and determining their numbers. In this contribution, we propose a test to verify whether a (one-way or two-way) matrix factor structure exists or not. To the best of our knowledge, this is the first work with a hypothesis testing procedure to discern between a genuine two-way matrix factor model (i.e. (1.2)), a one-way matrix factor structure (i.e. the first two cases of (1.10)), or no factors at all (i.e. the last case of (1.10)). Our procedures serve as a model checking tool to draw practical implications, e.g. on the estimation technique to be employed.

Formally, we develop tests for the following general hypotheses:

$$H_{i0} : k_i \geq k_i^0, \quad \text{vs.} \quad H_{i1} : k_i < k_i^0, \quad i = 1, 2, \tag{1.11}$$

where k_1^0 and k_2^0 are the hypothesised numbers of row and column factors, respectively. Our tests exploit the eigen-gap property of the second moment matrix of the matrix series: we show that if there are k_i^0 common row (or column) factors, then the largest k_i^0 eigenvalues diverge ALMOST SURELY, as the matrix dimensions increase, at a faster rate than the remaining ones. To the best of our knowledge, for the first time in the literature of matrix factor analysis, this paper obtains an almost-sure (not just in probability) diverging lower bound of the largest k_i^0 eigenvalues of the column (or row) covariance matrix with and without projection, and an almost-sure upper bound of the remaining eigenvalues. We then exploit the almost-sure eigen-gap, thereby constructing a randomised test in a similar manner to Trapani (2018). In order to avoid the non-reproducibility issue of randomised tests, we propose a "strong" rule to decide between H_{i0} and H_{i1}, inspired by the Law of the Iterated Logarithm.

Our approach has several desirable features. First, it is based on testing, and therefore it does not suffer from the arbitrariness in thresholding the eigenvalues, which is typical of information criteria. Second, it can also be used to test for $H_{i0} : k_i \geq 1$ versus $H_{i1} : k_i = 0$, thus avoiding the arbitrariness of having to create an "artificial" eigenvalue, which is typically used to initialise procedures based on eigenvalue ratios. Third, our tests – and therefore our decision rules – do not require (virtually) any restrictions on the relative rates of divergence of p_1, p_2 and T as they pass to infinity (apart from the very mild one in Assumption B5), nor do they require the white noise assumption on the idiosyncratic error matrix as in Wang et al. (2019). As far as the last point is concerned, we would like to mention that the set-up by Wang et al. (2019) (see also Lam and Yao, 2012) assumes that E_t is white noise, although, as a trade-off, less restrictive

assumptions are needed on the cross-sectional correlation among the components of E_t. In the context of such a set-up, the factor model can be validated by using existing high-dimensional white noise tests. Conversely, in the context of an approximate factor model like ours, the issue of model validation has not been fully investigated, i.e. no test exists to check that there is indeed a factor structure. Our paper fills the gap in literature, and, in general, is applicable to a wide variety of datasets.

In addition to diagnosing matrix structures, tests for (1.11) can be cast in a sequential procedure, as e.g. in Onatski (2009) and Trapani (2018), thereby obtaining an estimator for the number of common row (and/or column) factors. To the best of our knowledge, this is the first estimator of the numbers of row and/or column factors specifically designed for large matrix sequence, not based on eigenvalue thresholding. After determining the common factor dimensions, it is possible to apply the inferential theory developed e.g. in Chen and Fan (2021), or Yu et al. (2022). We propose two methodologies to test for (1.11), based on the eigenvalues of two different sample second moment matrices. Our first procedure is based on evaluating the k_i^0-th largest eigenvalues of the row (when $i = 1$) and column (when $i = 2$) "flattened" sample covariance matrices, defined as

$$M_c := \frac{1}{Tp_2}\sum_{t=1}^{T} X_t X_t' = \frac{1}{Tp_2}\sum_{t=1}^{T}\sum_{i=1}^{p_2} X_{\cdot i,t} X_{\cdot i,t}',$$
$$M_r := \frac{1}{Tp_1}\sum_{t=1}^{T} X_t' X_t = \frac{1}{Tp_1}\sum_{t=1}^{T}\sum_{j=1}^{p_1} X_{j\cdot,t} X_{j\cdot,t}',$$

where $X_{\cdot i,t}$ denotes the ith column of X_t, and $X_{j\cdot,t}$ its jth row. This testing procedure is computationally straightforward, and it requires only one step. On the other hand, using M_c and M_r ignores the two-way factor structure in model (1.2). Hence, we also propose a second, two-step methodology which makes full use of the low-rank structure of the common component matrix in (1.2). In particular, we test for (1.11) based on the column covariance matrix of a projected matrix time series, inspired by Yu et al. (2022).

We would like to point out that our set-up, despite its generality, still requires some restrictions on the data generating process of X_t. Indeed, whilst we allow for weak cross-sectional dependence among the idiosyncratic components, we would like to point out that the recent contribution by Lam (2021) considers a different, stronger form of dependence in the idiosyncratic errors, arising from the presence of weak common factors. Further, in our theory, we do not consider the presence of weak factors (see, however, the discussion in Section 3.4.2), which may be viewed as a shortcoming of our set-up; however, in Section 3.4.2, we briefly discuss this case, indicating that it can also be studied with our methodology. Also, our estimator of, say, k_1 based on the projection estimator of Yu et al. (2022) requires $k_2 > 0$, which therefore must be tested beforehand (see the discussion in Section B.3.4 in the supplement). Finally, a key requirement for our approach is that the specification in (1.2) and (1.10) is correct, i.e. that there is a Kronecker product structure in the loadings as indicated in (1.3); in the concluding section, we further discuss the implications of this assumption.

The rest of the paper is organised as follows. Section 2 presents the main assumptions and results on the spectra of M_c and M_r, as well as the projection-based second moment matrices. Section 3 gives two hypotheses testing procedures for (1.11), and the sequential testing methodology to determine k_i for $i = 1$ and 2; in particular, our "strong" rule to decide between H_{i0} and H_{i1} is given in Section 3.2. We evaluate our theory through an extensive simulation exercise in Section 4, and we further illustrate our findings through two empirical applications in Section 5. Section 6 concludes the paper and discusses some avenues for future research.

To end this section, we introduce some further notation in addition to the one already defined above. Positive finite constants are denoted as $c_0, c_1, \ldots$, and their values may change from line to line. Throughout the paper, we use the shorthand notation "a.s." for "almost sure(ly)". Given two sequences $a_{p_1,p_2,T}$ and $b_{p_1,p_2,T}$, we say that $a_{p_1,p_2,T} = o_{a.s.}\left(b_{p_1,p_2,T}\right)$ if, as $\min\{p_1, p_2, T\} \to \infty$, it holds that $a_{p_1,p_2,T} b_{p_1,p_2,T}^{-1} \to 0$ a.s.; we say that $a_{p_1,p_2,T} = O_{a.s.}\left(b_{p_1,p_2,T}\right)$ to denote that as $\min\{p_1, p_2, T\} \to \infty$, it holds that $a_{p_1,p_2,T} b_{p_1,p_2,T}^{-1} \to c_0 < \infty$ a.s.; and we use the notation $a_{p_1,p_2,T} = \Omega_{a.s.}\left(b_{p_1,p_2,T}\right)$ to indicate that as $\min\{p_1, p_2, T\} \to \infty$, it holds that $a_{p_1,p_2,T} b_{p_1,p_2,T}^{-1} \to c_0 > 0$ a.s. Given an $m \times n$ matrix A, we denote its transpose as A' and its element in position $\{i, j\}$ as A_{ij} or a_{ij}, i.e. using either upper or lower case letters. Further, we denote the spectral norm as $\|A\|$; we use $\|A\|_{\max}$ to denote the maximum of the absolute values of A's elements; finally, we let $\lambda_i(A)$ be the ith largest eigenvalue of A. Other, relevant notation is introduced later on in the paper.

2. Spectra

We study the eigenvalues of the covariance matrices M_c and M_r, and of the projected versions (denoted as $\widetilde{M}_c$ and $\widetilde{M}_r$). In both cases, we find that the matrices have an eigen-gap between the first k_1 (resp. k_2) eigenvalues and the remaining ones. As the cross-sectional sample size p_1 (resp. p_2), increases, the first k_1 (resp. k_2) eigenvalues diverge at a faster rate than the remaining ones.

2.1. Assumptions

The following assumptions are borrowed from the paper by Yu et al. (2022), to which we refer for detailed explanations. In Section A of the Supplementary Material, we discuss some of our assumptions in greater detail.

Assumption B1. (i) (a) $E(F_t)=0$, and (b) $E\|F_t\|^{4+\epsilon}\leq c_0$, for some $\epsilon>0$; (ii) when $k_i>0$ for $i=1,2$, it holds that

$$\frac{1}{T}\sum_{t=1}^{T}F_tF_t'\overset{a.s.}{\to}\Sigma_1 \text{ and } \frac{1}{T}\sum_{t=1}^{T}F_t'F_t\overset{a.s.}{\to}\Sigma_2, \tag{2.1}$$

where Σ_i is a $k_i\times k_i$ positive definite matrix with distinct eigenvalues, $\lambda_{\max}(\Sigma_i)<\infty$, and spectral decomposition $\Sigma_i=\Gamma_i\Lambda_i\Gamma_i'$. The factor numbers k_1 and k_2 are fixed as $\min\{T,p_1,p_2\}\to\infty$; (iii) it holds that, for all h_1,l_1 and h_2,l_2

$$E\max_{1\leq\tilde{t}\leq T}\left(\sum_{t=1}^{\tilde{t}}\left(F_{h_1h_2,t}F_{l_1l_2,t}-E\left(F_{h_1h_2,t}F_{l_1l_2,t}\right)\right)\right)^2\leq c_0T;$$

(iv) (a) when $k_2=0$ and $k_1>0$, it holds that

$$\lambda_{\max}\left(\frac{1}{T}\sum_{t=1}^{T}F_t'F_t\right)=O_{a.s.}\left(\left(1+\sqrt{\frac{p_2}{T}}\right)^2\right) \text{ and } \frac{1}{Tp_2}\sum_{t=1}^{T}F_tF_t'\overset{a.s.}{\to}\Sigma_1^*,$$

with Σ_1^* a $k_1\times k_1$ positive definite matrix with distinct eigenvalues and $\lambda_{\max}\left(\Sigma_1^*\right)<\infty$; (b) when $k_1=0$ and $k_2>0$, it holds that

$$\lambda_{\max}\left(\frac{1}{T}\sum_{t=1}^{T}F_tF_t'\right)=O_{a.s.}\left(\left(1+\sqrt{\frac{p_1}{T}}\right)^2\right) \text{ and } \frac{1}{Tp_1}\sum_{t=1}^{T}F_t'F_t\overset{a.s.}{\to}\Sigma_2^*,$$

with Σ_2^* a $k_2\times k_2$ positive definite matrix with distinct eigenvalues and $\lambda_{\max}\left(\Sigma_2^*\right)<\infty$.

Assumption B2. (i) $\|R\|_{\max}\leq c_0$, and $\|C\|_{\max}\leq c_1$; (ii) as $\min\{p_1,p_2\}\to\infty$, $\|p_1^{-1}R'R-I_{k_1}\|\to 0$ and $\|p_2^{-1}C'C-I_{k_2}\|\to 0$.

Assumptions B1 and B2 are standard in large factor models, and we refer, for example, to Chen and Fan (2021). In Assumption B1*(i)*(b), note the (mild) strengthening of the customarily assumed fourth moment existence condition on F_t — this is required in order to prove our results, which rely on almost sure rates. Similarly, the maximal inequality in part *(iii)* of the assumption is usually not considered in the literature, and it can be derived from more primitive dependence assumptions: for example, it can be shown to hold under various mixing conditions (see e.g. Rio, 1995; and Shao, 1995); in Section A.2 in the Supplementary Material, we show its validity for the very general class of decomposable Bernoulli shifts (see e.g. Wu, 2005). Part *(iv)* of the assumption is needed to study the case where $k_i=0$ — in that case, according to (1.10), F_t is "large" along one dimension. The bound on $\lambda_{\max}\left(T^{-1}\sum_{t=1}^{T}F_t'F_t\right)$ is a high-level condition, which we borrow from the literature on large Random Matrix Theory (RMT; see the seminal paper by Geman, 1980, and the review in El Karoui, 2005). In Section A.3 in the Supplementary Material, we also discuss what happens under more primitive assumptions which do not require the use of RMT.

Finally, we point out that, according to Assumption B2, the common factors are pervasive. Extensions to the case of "weak" factors – where the norms of R and C diverge at a slower rate than $p_1^{1/2}$ and $p_2^{1/2}$ – are briefly discussed in Section 3.4.2.

Assumption B3. (i) (a) $E(e_{ij,t})=0$, and (b) $E\left|e_{ij,t}\right|^8\leq c_0$; (ii) for all $1\leq t\leq T$, $1\leq i\leq p_1$ and $1\leq j\leq p_2$,

$$\text{(a).}\sum_{s=1}^{T}\sum_{l=1}^{p_1}\sum_{h=1}^{p_2}|E(e_{ij,t}e_{lh,s})|\leq c_0,\quad \text{(b).}\sum_{l=1}^{p_1}\sum_{h=1}^{p_2}|E(e_{lj,t}e_{ih,t})|\leq c_0;$$

(iii) for all $1\leq t\leq T$, $1\leq i,l_1\leq p_1$ and $1\leq j,h_1\leq p_2$,

$$\begin{aligned}
&\text{(a)}.\quad \textstyle\sum_{s=1}^{T}\sum_{l_2=1}^{p_1}\sum_{h=1}^{p_2}\left|Cov(e_{ij,t}e_{l_1j,t},e_{ih,s}e_{l_2h,s})\right|\leq c_0,\\
&\qquad\quad \textstyle\sum_{s=1}^{T}\sum_{l=1}^{p_1}\sum_{h_2=1}^{p_2}\left|Cov(e_{ij,t}e_{ih_1,t},e_{lj,s}e_{lh_2,s})\right|\leq c_0,\\
&\qquad\quad \textstyle\sum_{s=1}^{T}\sum_{l=1}^{p_1}\sum_{h=1}^{p_2}\left|Cov(e_{ij,t}^2,e_{lh,s}^2)\right|\leq c_0,\\
&\text{(b)}.\quad \textstyle\sum_{s=1}^{T}\sum_{l_2=1}^{p_1}\sum_{h_2=1}^{p_2}\left|Cov(e_{ij,t}e_{l_1h_1,t},e_{ij,s}e_{l_2h_2,s})+Cov(e_{l_1j,t}e_{ih_1,t},e_{l_2j,s}e_{ih_2,s})\right|\leq c_0,
\end{aligned}$$

(iv) it holds that $\lambda_{\min}\left[E\left(\frac{1}{p_2T}\sum_{t=1}^{T}E_tE_t'\right)\right]>0$ and $\lambda_{\min}\left[E\left(\frac{1}{p_1T}\sum_{t=1}^{T}E_t'E_t\right)\right]>0$.

Assumption B3 ensures the (cross-sectional and time series) summability of the idiosyncratic terms E_t. The assumption allows for (weak) dependence in both the space and time domains, and – as also mentioned in the introduction – it can be read in conjunction with the paper by Wang et al. (2019), where E_t is assumed to be white noise, but no structure is assumed on its covariance matrix. In Section A.2 in the Supplementary Material, we show that the time-series properties of E_t (in particular parts *(ii)* and *(iii)*, which are high-level assumptions) are satisfied, similarly to Assumption B1 , by the wide class of decomposable Bernoulli shifts.

Assumption B4. (i) For any deterministic vectors v and w satisfying $\|v\| = 1$ and $\|w\| = 1$ with suitable dimensions,

$$E\left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} F_t (v' E_t w) \right\|^2 \leq c_0;$$

(ii) for all $1 \leq i, l_1 \leq p_1$ and $1 \leq j, h_1 \leq p_2$,

$$\text{(a).} \left\| \sum_{h=1}^{p_2} E(\bar{\zeta}_{ij} \otimes \bar{\zeta}_{ih}) \right\|_{\max} \leq c_0, \quad \left\| \sum_{l=1}^{p_1} E(\bar{\zeta}_{ij} \otimes \bar{\zeta}_{lj}) \right\|_{\max} \leq c_0,$$

$$\text{(b).} \left\| \sum_{l=1}^{p_1} \sum_{h_2=1}^{p_2} Cov(\bar{\zeta}_{ij} \otimes \bar{\zeta}_{ih_1}, \bar{\zeta}_{lj} \otimes \bar{\zeta}_{lh_2}) \right\|_{\max} \leq c_0, \left\| \sum_{l_2=1}^{p_1} \sum_{h=1}^{p_2} Cov(\bar{\zeta}_{ij} \otimes \bar{\zeta}_{l_1 j}, \bar{\zeta}_{ih} \otimes \bar{\zeta}_{l_2 h}) \right\|_{\max} \leq c_0,$$

where $\bar{\zeta}_{ij} = \text{Vec}(\sum_{t=1}^{T} F_t e_{ij,t}/\sqrt{T})$; (iii) (a) when $k_2 = 0$, it holds that

$$\max_{1 \leq h, h' \leq k_2} \sum_{j=1}^{p_2} \sum_{t=1}^{T} \left| E\left(F_{ih,t} F_{i'h',s} e_{ij,t} e_{i'j',s} \right) \right| \leq c_0,$$

for all $i, i', j' \neq j$ and $s \neq t$; (b) when $k_1 = 0$, it holds that

$$\max_{1 \leq h, h' \leq k_1} \sum_{j=1}^{p_1} \sum_{t=1}^{T} \left| E\left(F_{hi,t} F_{h'i',s} e_{ij,t} e_{i'j',s} \right) \right| \leq c_0,$$

for all $i, i', j' \neq j$ and $s \neq t$.

According to Assumption B4, the common factors F_t and the errors E_t can be weakly correlated. Part *(i)* of the assumption is similar to e.g. Assumption D in Bai (2003), in the context of vector factor models, and it is easy to see that it is satisfied e.g. when $\{F_t\}$ and $\{E_t\}$ are two mutually independent groups. As far as part *(ii)* is concerned, this is a more high-level assumption which is required in order for Lemma B.3 in Yu et al. (2022) to hold; in turn, this ensures that the "initial" estimators of R and C required in Section 2.3 are consistent, also providing a rate for them. The assumption is similar, in spirit, to Assumption 4(3) in Chen and Fan (2021), and to Assumption D in Bai (2003), in the case of vector valued series. In Section A.4 in the Supplementary Material, we discuss some cases in which this part of Assumption B4 is satisfied, including the case where $e_{ij,t} = g(F_t) u_{ij,t}$, with $u_{ij,t}$ independent across i, j, and t, and $g(\cdot)$ a measurable function; again, a similar case is also mentioned in the discussion of Assumption D in Bai (2003).

Assumption B5. There exist $\kappa_1, \kappa_2 < \infty$ such that

$$p_2, T = O\left(p_1^{\kappa_1}\right) \quad \text{and} \quad p_1, T = O\left(p_2^{\kappa_2}\right).$$

By Assumption B5, the relative rates at which the sample sizes p_1, p_2 and T pass to infinity can be very large, but not "too large" – i.e., cases like $p_1 = \exp(p_2)$ are ruled out. This restriction is arguably very mild; furthermore, it is owed to our method of proof, and we conjecture that it could be lifted, allowing also for extreme cases such as $p_1 = \exp(p_2)$. From an operational standpoint, the restriction is going to be satisfied by virtually all datasets.

2.2. The spectra of M_c and M_r

To avoid repetitions, we only present results for M_c; the spectrum of M_r can be studied exactly in the same way. We use the short-hand notation λ_j to indicate the jth largest eigenvalue of the expectation of M_c, and use $\widehat{\lambda}_j$ denote the jth largest eigenvalue of M_c.

Our first theorem provides an a.s. eigen-gap for M_c.

Theorem 1. *Suppose that Assumptions B1–B4 are satisfied. When $k_1 > 0$, it holds that*

$$\widehat{\lambda}_j = \Omega_{a.s.}(p_1), \tag{2.2}$$

for all $j \leq k_1$; further, there exist a constant $c_0 < \infty$ such that

$$\widehat{\lambda}_j = c_0 + o_{a.s.}\left(\frac{p_1}{\sqrt{Tp_2}}\left(\ln^2 p_1 \ln p_2 \ln T\right)^{1/2+\epsilon}\right), \tag{2.3}$$

for all $j > k_1$, and all $\epsilon > 0$. When $k_1 = 0$, it holds that

$$\widehat{\lambda}_j = c_0 + O_{a.s.}\left(\frac{p_1}{T}\right) + o_{a.s.}\left(\frac{p_1}{\sqrt{Tp_2}}\left(\ln^2 p_1 \ln p_2 \ln T\right)^{1/2+\epsilon}\right), \tag{2.4}$$

for all $j \geq 1$, and all $\epsilon > 0$.

The eigen-gap in the spectrum of M_c is the building block to construct a procedure to decide between H_{i0} and H_{i1} in (1.11). We point out that, although the results in (2.2) and (2.3) are similar, in spirit, to the ones derived by Trapani (2018), here we follow a quite different method of proof. Using the approach in Trapani (2018), we would be able to show only the rate $o_{a.s.}\left(T^{-1/2}p_1\left(\ln^2 p_1 \ln p_2 \ln T\right)^{1/2+\epsilon}\right)$ in (2.3), thus having a (much) worse rate; moreover, the case $k_i = 0$, where F_t has growing dimension, is not covered by Trapani (2018). As far as (2.4) is concerned, we note that this is a consequence of Assumption B1 *(iv)*, and in particular of the bound $\lambda_{\max}\left(T^{-1}\sum_{t=1}^{T} F_t F_t'\right) = O_{a.s.}\left(\left(1+\sqrt{p_1/T}\right)^2\right)$ – see also Section A.3 in the Supplementary Material for a discussion. Finally, we would like to point out that the theorem holds for any value of k_2, including $k_2 = 0$.

2.3. The spectra of projected covariance matrices

The matrices M_c and M_r are straightforward to compute and use, but they are based on the implicit assumption that a "one-way" factor structure is present only in the columns (or rows) of the observations.

When $k_2 > 0$, we propose to fully make use of the two-way interactive factor structure in (1.2), by studying the spectrum of a projected column (row) covariance matrix, as suggested by Yu et al. (2022). Heuristically, if $k_2 > 0$ and if C is known and satisfies the orthogonality condition $C'C/p_2 = I_{k_2}$, the data matrix can be projected into a lower dimensional space by setting $Y_t = X_t C/p_2$. In view of this, we define

$$\widetilde{M}_c = \frac{1}{T}\sum_{t=1}^{T} \widetilde{Y}_t \widetilde{Y}_t',$$

where $\widetilde{Y}_t = X_t \widehat{C}/p_2$ and $\widehat{C}$ is an initial estimator of C ($\widetilde{M}_r$ can be defined similarly). As suggested by Yu et al. (2022), the initial estimator can be set as $\widehat{C} = \sqrt{p_2}Q$, where the columns of Q are the leading k_2 eigenvectors of M_r.

Let $\widetilde{\lambda}_j$ denote the jth largest eigenvalue of $\widetilde{M}_c$. The following result measures the eigen-gap of $\widetilde{M}_c$.

Theorem 2. *We assume that Assumption B1–Assumption B4 are satisfied and that $k_2 > 0$. When $k_1 > 0$, it holds that*

$$\widetilde{\lambda}_j = \Omega_{a.s.}(p_1), \tag{2.5}$$

for all $j \leq k_1$; further, it holds that

$$\widetilde{\lambda}_j = o_{a.s.}\left(\left(\frac{1}{p_2} + \frac{1}{T} + \frac{p_1}{\sqrt{Tp_2}}\right)\left(\ln^2 p_1 \ln p_2 \ln T\right)^{1+\epsilon}\right), \tag{2.6}$$

for all $j > k_1$ and all $\epsilon > 0$. When $k_1 = 0$, it holds that

$$\widetilde{\lambda}_j = O_{a.s.}\left(\frac{p_1}{T}\right) + o_{a.s.}\left(\left(\frac{1}{p_2} + \frac{1}{T} + \frac{p_1}{\sqrt{Tp_2}}\right)\left(\ln^2 p_1 \ln p_2 \ln T\right)^{1+\epsilon}\right), \tag{2.7}$$

for all $j \geq 1$, and all $\epsilon > 0$.

Comparing (2.6) with (2.3) in Theorem 1, the eigen-gap of $\widetilde{M}_c$ is wider than that of M_c. Thus, using $\widetilde{M}_c$ should yield a higher testing power and a better estimate of k_1 (and/or k_2) if the two-way interactive factor structure is really true in practice. Of course, this is predicated upon having $k_2 > 0$. As also mentioned after Theorem 1, the rate in (2.6) is sharper than one would find following method of proof in Trapani (2018); even in this case, we would only obtain the rate $o_{a.s.}\left(T^{-1/2}p_1\left(\ln^2 p_1 \ln p_2 \ln T\right)^{1+\epsilon}\right)$, which again would be sub-optimal. Finally, the case where $k_1 = 0$ and $k_2 > 0$ is covered by Eq. (2.7): the same comments as for (2.4) apply in this case (see also Section A.3).

3. Inference on the number of factors

In this section, we investigate two related problems, based on determining the dimension of the (row or column) factor structures. For brevity, we only report results concerning k_1, but all our procedures can be readily extended to analyse k_2.

We begin by presenting the tests for the null that $H_0 : k_1 \geq k_1^0$ for a given k_1^0 (we omit the subscript i in H_{i0} for simplicity). We then apply these to determining whether there is a factor structure; if this is the case, we develop a sequential procedure to determine the dimension of each factor space. Both procedures are based on constructing, as a first step, a test based on the rates of divergence of the eigenvalues of either M_c or $\widetilde{M}_c$ (see Section 3.1); and, as a second step, a decision rule to choose between H_0 and H_1 which is not affected by the randomness added by the researcher (Section 3.2).

3.1. Hypothesis testing and the randomised tests

We consider tests for

$$H_0 : k_1 \geq k_1^0 \text{ vs. } H_1 : k_1 < k_1^0 \quad \text{for some } k_1^0 \in [1, \dots, k_{\max}], \tag{3.1}$$

where $k_{\max}$ is a pre-specified upper bound. The hypothesis in (3.1) is equivalent to the following hypothesis on the eigenvalue $\lambda_{k_1^0}$, that is

$$H_0 : \lambda_{k_1^0} \geq c_0 p_1 \text{ vs. } H_1 : \lambda_{k_1^0} \leq c_0. \tag{3.2}$$

We propose two types of test statistics for the hypothesis testing problem in (3.2). Let $\beta = \ln p_1 / \ln(p_2 T)$, and let $\delta = \delta(\beta) \in (0, 1)$, such that

$$\begin{cases} \delta = \varepsilon & \text{if } \beta \leq 1/2 \\ \delta = 1 - 1/(2\beta) + \varepsilon & \text{if } \beta > 1/2 \end{cases}, \tag{3.3}$$

where $\varepsilon > 0$ is an arbitrarily small, user-defined number.[3] Given δ, we define

$$\widehat{\phi}_{k_1^0} = \exp\left\{ \frac{p_1^{-\delta} \widehat{\lambda}_{k_1^0}}{p_1^{-1} \sum_{j=1}^{p_1} \widehat{\lambda}_j} \right\} - 1 \text{ and } \widetilde{\phi}_{k_1^0} = \exp\left\{ \frac{p_1^{-\delta} \widetilde{\lambda}_{k_1^0}}{p_1^{-1} \sum_{j=1}^{p_1} \widetilde{\lambda}_j} \right\} - 1; \tag{3.4}$$

these are transformations of $\widehat{\lambda}_{k_1^0}$ and $\widetilde{\lambda}_{k_1^0}$, rescaled by the trace of M_c and $\widetilde{M}_c$ respectively, to make them scale-invariant.

The choice of δ in (3.3) is an important specification. Its purpose is the same as in Trapani (2018), i.e. to make $p_1^{-\delta}\widehat{\lambda}_{k_1^0}$ (and $p_1^{-\delta}\widetilde{\lambda}_{k_1^0}$) drift to zero when $\lambda_{k_1^0} \leq c_0$. In this case, it is easy to see that rescaling by $p_1^{-\delta}$ gets rid of the estimation error. Indeed, as an example, by Theorem 2

$$p_1^{-\delta}\widetilde{\lambda}_{k_1^0} = o_{a.s.}\left(\left(\frac{p_1^{-\delta}}{p_2} + \frac{p_1^{-\delta}}{T} + \frac{p_1^{1-\delta}}{\sqrt{Tp_2}} \right) \left(\ln^2 p_1 \ln p_2 \ln T \right)^{1+\epsilon} \right);$$

clearly, $p_2^{-1} p_1^{-\delta} + T^{-1} p_1^{-\delta} = o(1)$. Also, whenever $\beta \leq 1/2$, it is immediate to see that $\ln p_1 / \ln(p_2 T) \leq 1/2 \Rightarrow p_1 = O\left((p_2 T)^{1/2}\right)$, so that

$$\frac{p_1^{1-\delta}}{\sqrt{Tp_2}} = \frac{p_1^{1-\varepsilon}}{\sqrt{Tp_2}} = o\left(p_1^{-\varepsilon}\right),$$

and therefore, on account of Assumption B5

$$\begin{aligned} p_1^{-\delta}\widetilde{\lambda}_{k_1^0} &= o_{a.s.}\left(\left(\frac{p_1^{-\delta}}{p_2} + \frac{p_1^{-\delta}}{T} + \frac{p_1^{1-\delta}}{\sqrt{Tp_2}} \right) \left(\ln^2 p_1 \ln p_2 \ln T \right)^{1+\epsilon} \right) \\ &= o_{a.s.}\left(p_1^{-\varepsilon} \left(\ln^2 p_1 \ln p_2 \ln T \right)^{1+\epsilon} \right) = o_{a.s.}(1). \end{aligned}$$

On the other hand, when $\beta > 1/2$, using the definition of β

$$\begin{aligned} \frac{p_1^{1-\delta}}{\sqrt{Tp_2}} &= \frac{p_1^{1/(2\beta)-\varepsilon}}{\sqrt{Tp_2}} = \exp\left(\left(\frac{1}{2\beta} - \varepsilon \right) \ln p_1 - \frac{1}{2} \ln(p_2 T) \right) \\ &= \exp\left(\left(\frac{\ln(p_2 T)}{2 \ln p_1} - \varepsilon \right) \ln p_1 - \frac{1}{2} \ln(p_2 T) \right) \\ &= \exp(-\varepsilon \ln p_1) = p_1^{-\varepsilon}, \end{aligned}$$

whence the same passages as above yield the desired result. Conversely, $\delta < 1$, thus still allowing $p_1^{-\delta}\widehat{\lambda}_{k_1^0}$ (and $p_1^{-\delta}\widetilde{\lambda}_{k_1^0}$) to pass to infinity if $\lambda_{k_1^0}$ does diverge. It can be verified that the value of δ in (3.3) suffices to make the estimation error drift

[3] Our simulations show that results are not overly sensitive to the choice of ε, and that using $\varepsilon = 0.01$ yields good results for all the scenario considered. We therefore recommend this choice as guideline.

to zero in all cases covered by Theorems 1 and 2, with the exception of the cases covered by (2.4) and (2.7), i.e. when testing for $H_0 : k_1 \geq 1$, versus $H_1 : k_1 = 0$. In that case, based on the rates in (2.4) and (2.7), it can be shown that (3.3) can be employed after replacing β with $\beta' = \ln p_1 / \min\{\ln(p_2 T), 2\ln(T)\}$.

We now turn to discussing how to use $\widehat{\phi}_{k_1^0}$ and $\widetilde{\phi}_{k_1^0}$ to test for $H_0 : k_1 \geq k_1^0$ in (3.2). Theorems 1 and 2 provide rates for $\widehat{\lambda}_{k_1^0}$ and $\widetilde{\lambda}_{k_1^0}$ (and, consequently, also for $\widehat{\phi}_{k_1^0}$ and $\widetilde{\phi}_{k_1^0}$) under both the null and the alternative in the hypothesis testing framework in (3.2), but no limiting distribution is available. Hence, we propose to randomise $\widehat{\phi}_{k_1^0}$ and $\widetilde{\phi}_{k_1^0}$, in a similar way to Trapani (2018):

Step 1 Generate *i.i.d.* samples $\{\eta^{(m)}\}_{m=1}^M$ with common distribution $N(0, 1)$.

Step 2 Given $\{\eta^{(m)}\}_{m=1}^M$, construct sample sets $\{\widehat{\psi}_{k_1^0}^{(m)}(u)\}_{m=1}^M$ and $\{\widetilde{\psi}_{k_1^0}^{(m)}(u)\}_{m=1}^M$ as

$$\widehat{\psi}_{k_1^0}^{(m)}(u) = I\left[\sqrt{\widehat{\phi}_{k_1^0}} \times \eta^{(m)} \leq u\right], \quad \widetilde{\psi}_{k_1^0}^{(m)}(u) = I\left[\sqrt{\widetilde{\phi}_{k_1^0}} \times \eta^{(m)} \leq u\right].$$

Step 3 Define

$$\widehat{\nu}_{k_1^0}(u) = \frac{2}{\sqrt{M}} \sum_{m=1}^{M} \left[\widehat{\psi}_{k_1^0}^{(m)}(u) - \frac{1}{2}\right], \quad \widetilde{\nu}_{k_1^0}(u) = \frac{2}{\sqrt{M}} \sum_{m=1}^{M} \left[\widetilde{\psi}_{k_1^0}^{(m)}(u) - \frac{1}{2}\right]. \tag{3.5}$$

Step 4 The test statistics are finally defined as

$$\widehat{\Psi}_{k_1^0} = \int_U \left[\widehat{\nu}_{k_1^0}(u)\right]^2 dF(u), \quad \widetilde{\Psi}_{k_1^0} = \int_U \left[\widetilde{\nu}_{k_1^0}(u)\right]^2 dF(u),$$

where $F(u)$ is a weight function.

The test described above is similar to the one proposed in Trapani (2018); however, in the construction of $\widehat{\Psi}_{k_1^0}$ and $\widetilde{\Psi}_{k_1^0}$, we propose a weighted average across different values of u through the weight function $F(u)$. As a consequence, it can be expected that the test will not be affected by an individual value of u, a form of scale invariance which is not considered in Trapani (2018).

Assumption C1. $F(u)$ is a differentiable function for all $u \in U$ such that: (i) $\int_U dF(u) = 1$; (ii) $\int_U u^2 dF(u) < \infty$.

Assumption C1 is satisfied by several functions, the most "natural" candidates being distribution functions; we discuss in detail $F(u)$, its possible specifications, and how to compute integrals involving it in Section B.2 of the Supplementary Material.

Let P^* denote the probability law of $\{\widehat{\psi}_{k_1^0}^{(m)}(u)\}_{m=1}^M$ and $\{\widetilde{\psi}_{k_1^0}^{(m)}(u)\}_{m=1}^M$ conditional on the sample $\{X_t, 1 \leq t \leq T\}$, and "$\xrightarrow{D^*}$" and "$\xrightarrow{P^*}$" as convergence in distribution and in probability, respectively, according to P^*.

Proposition 1. *We assume that Assumptions B1–B5 and C1 are satisfied. Then, under $H_0 : k_1 \geq k_1^0$, as $\min\{p_1, p_2, T, M\} \to \infty$ with*

$$M \exp\left\{-\epsilon p_1^{1-\delta}\right\} \to 0, \tag{3.6}$$

for some $0 < \epsilon < c_0/\bar{\lambda}$ and $\bar{\lambda} = p_1^{-1} \sum_{j=1}^{p_1} \lambda_j$, it holds that

$$\widehat{\Psi}_{k_1^0} \xrightarrow{D^*} \chi_1^2, \tag{3.7}$$

for almost all realisations of $\{X_t, 1 \leq t \leq T\}$. Under the same assumptions, if $k_2 > 0$ it also holds that $\widetilde{\Psi}_{k_1^0} \xrightarrow{D^*} \chi_1^2$ for almost all realisations of $\{X_t, 1 \leq t \leq T\}$.

Under $H_1 : \lambda_{k_1^0} \leq c_0 < \infty$, as $\min\{p_1, p_2, T, M\} \to \infty$ it holds that

$$M^{-1}\widehat{\Psi}_{k_1^0} \xrightarrow{P^*} c_1, \tag{3.8}$$

for some $0 < c_1 < \infty$ and almost all realisations of $\{X_t, 1 \leq t \leq T\}$. Under the same assumptions, if $k_2 > 0$ it also holds that $M^{-1}\widetilde{\Psi}_{k_1^0} \xrightarrow{P^*} c_1$, for almost all realisations of $\{X_t, 1 \leq t \leq T\}$.

Eq. (3.7) states that, under the null, both test statistics $\widehat{\Psi}_{k_1^0}$ and $\widetilde{\Psi}_{k_1^0}$ converge in distribution to a chi-square distribution with one degree of freedom. This can be understood heuristically by noting that, under the null, both $\widehat{\phi}_{k_1^0}$ and $\widetilde{\phi}_{k_1^0}$ go to

infinity, and therefore the variances of $\sqrt{\widehat{\phi}_{k_1^0}} \times \eta^{(m)}$ and $\sqrt{\widetilde{\phi}_{k_1^0}} \times \eta^{(m)}$ also pass to infinity. Thus, heuristically, $\{\widehat{\psi}_{k_1^0}^{(m)}(u)\}_{m=1}^M$ and $\{\widetilde{\psi}_{k_1^0}^{(m)}(u)\}_{m=1}^M$ follow a Bernoulli distribution with success probability 1/2. By the Central Limit Theorem, in (3.5) as M goes to infinity, both $\widehat{\nu}_{k_1^0}(u)$ and $\widetilde{\nu}_{k_1^0}(u)$ follow the standard normal distribution $N(0, 1)$ (conditional on the sample) asymptotically. The results hold for all samples, save for a zero measure set. By Proposition 1, it follows immediately that, for almost all realisations of $\{X_t, 1 \leq t \leq T\}$

$$\lim_{\min\{p_1,p_2,T,M\}\to\infty} P^*\left(\widehat{\Psi}_{k_1^0} > c_\alpha | H_0\right) = \alpha, \tag{3.9}$$

where c_α is such that $P\left(\chi_1^2 > c_\alpha\right) = \alpha$, and

$$\lim_{\min\{p_1,p_2,T,M\}\to\infty} P^*\left(\widehat{\Psi}_{k_1^0} > c_\alpha | H_1\right) = 1. \tag{3.10}$$

The results also hold if one substitutes $\widehat{\Psi}_{k_1^0}$ with $\widetilde{\Psi}_{k_1^0}$ in (3.9) and (3.10).

3.2. A "strong" rule to decide between H_0 and H_1

The tests are constructed by using added randomness, $\{\eta^{(m)}\}_{m=1}^M$, whose effect does not vanish asymptotically as would be the case e.g. when using the bootstrap. In turn, this entails that the properties of tests based on $\widehat{\Psi}_{k_1^0}$ (and $\widetilde{\Psi}_{k_1^0}$) are different from the properties of "standard" tests. Indeed, Eq. (3.10) has the classical interpretation: whenever a researcher uses $\widehat{\Psi}_{k_1^0}$ (and $\widetilde{\Psi}_{k_1^0}$), (s)he will reject the null, when false, with probability one. Conversely, the implications of (3.9) are subtler. Due to the artificial randomness $\{\eta^{(m)}\}_{m=1}^M$, different researchers using the same data will obtain different values of $\widehat{\Psi}_{k_1^0}$ and $\widetilde{\Psi}_{k_1^0}$, and, consequently, different p-values; indeed, if an infinite number of researchers were to carry out the test, the p-values would follow a uniform distribution on [0, 1]. Corradi and Swanson (2006) provide an alternative explanation, writing that "[...] as the sample size gets larger, all researchers always reject the null when false, while α% of the researchers always reject the null when it is true".

In order to address this problem, we propose a further step which, in essence, "de-randomizes" $\widehat{\Psi}_{k_1^0}$ and $\widetilde{\Psi}_{k_1^0}$. Each researcher, instead of computing $\widehat{\Psi}_{k_1^0}$ or $\widetilde{\Psi}_{k_1^0}$ just once, will compute the test statistic S times, at each iteration s generating a statistic $\widehat{\Psi}_{k_1^0,s}$ (or $\widetilde{\Psi}_{k_1^0,s}$) using a random sequence $\left\{\eta_s^{(m)}, 1 \leq m \leq M\right\}$, independent across $1 \leq s \leq S$, and thence defining, for some $\alpha \in (0, 1)$

$$\widehat{Q}_{k_1^0}(\alpha) = S^{-1} \sum_{s=1}^{S} I\left[\widehat{\Psi}_{k_1^0,s} \leq c_\alpha\right], \tag{3.11}$$

and the same when using $\widetilde{\Psi}_{k_1^0,s}$ — in this case obtaining $\widetilde{Q}_{k_1^0}(\alpha)$. A consequence of Proposition 1 is

$$\begin{aligned} &\lim_{\min(p_1,p_2,T,M,S)\to\infty} P^*\{\widehat{Q}_{k_1^0}(\alpha) = 1-\alpha\} = 1 \quad \text{for } H_0 : k_1 \geq k_1^0, \\ &\lim_{\min(p_1,p_2,T,M,S)\to\infty} P^*\{\widehat{Q}_{k_1^0}(\alpha) = 0\} = 1 \quad \text{for } H_1 : k_1 < k_1^0. \end{aligned} \tag{3.12}$$

Eq. (3.12) stipulates that, as $S \to \infty$, averaging across s in (3.11) washes out the added randomness in $\widehat{Q}_{k_1^0}(\alpha)$: all researchers using this procedure will obtain the same value of $\widehat{Q}_{k_1^0}(\alpha)$, thereby ensuring reproducibility. The function $\widehat{Q}_{k_1^0}(\alpha)$ corresponds to (the complement to one of) the "fuzzy decision", or "abstract randomised decision rule" reported in Eq. (1.1a) in Geyer and Meeden (2005a). Geyer and Meeden (2005a) (see also Geyer and Meeden, 2005b) provide a helpful discussion of the meaning of $\widehat{Q}_{k_1^0}(\alpha)$: the problem of deciding in favour or against H_0 may be modelled through a random variable, say D, which can take two values, namely "do not reject H_0 " and "reject H_0". Such a random variable has probability $\widehat{Q}_{k_1^0}(\alpha)$ to take the value "do not reject H_0 ", and probability $1 - \widehat{Q}_{k_1^0}(\alpha)$ to take the value "reject H_0 ". In this context, (3.12) states that (asymptotically), the probability of the event $\{\omega : D =$ "reject H_0"$\}$ is α when H_0 is satisfied, for all researchers — corresponding to the notion of *size* of a test; see also the quote from Corradi and Swanson (2006) reported above. Conversely, under H_1, the probability of the event $\{\omega : D =$ "reject H_0"$\}$ is 1 (asymptotically), corresponding to the notion of *power*.

Reporting the value of $\widehat{Q}_{k_1^0}(\alpha)$ or $\widetilde{Q}_{k_1^0}(\alpha)$ could be sufficient in some applications. In our case, the individual tests for $H_0 : k_1 \geq k_1^0$ will form the basis of a sequential procedure to provide an estimate of k_1, and therefore we also need a decision rule to choose, based on $\widehat{Q}_{k_1^0}(\alpha)$ (or $\widetilde{Q}_{k_1^0}(\alpha)$), between H_0 and H_1. We base such a decision rule on a Law of the Iterated Logarithm for $\widehat{Q}_{k_1^0}(\alpha)$ and $\widetilde{Q}_{k_1^0}(\alpha)$.

Theorem 3. *We assume that Assumptions B1–B5 and C1 are satisfied, and that $M = O(T)$ and $S = \Omega(M)$. Then it holds that*

$$\frac{\widehat{Q}_{k_1^0}(\alpha) - (1-\alpha)}{\sqrt{\alpha(1-\alpha)}} = \Omega_{a.s.}\left(\sqrt{\frac{2\ln\ln S}{S}}\right), \quad \frac{\widetilde{Q}_{k_1^0}(\alpha) - (1-\alpha)}{\sqrt{\alpha(1-\alpha)}} = \Omega_{a.s.}\left(\sqrt{\frac{2\ln\ln S}{S}}\right), \tag{3.13}$$

under $H_0 : k_1 \geq k_1^0$, for almost all realisations of $\{X_t, 1 \leq t \leq T\}$. Also, it holds that

$$\widehat{Q}_{k_1^0}(\alpha) = o_{a.s.}(1) \quad \text{and} \quad \widetilde{Q}_{k_1^0}(\alpha) = o_{a.s.}(1), \tag{3.14}$$

under $H_1 : k_1 < k_1^0$, for almost all realisations of $\{X_t, 1 \leq t \leq T\}$.

Eqs. (3.13) and (3.14) complement (3.12), and quantify the gap in the asymptotic behaviour of $\widehat{Q}_{k_1^0}(\alpha)$ (or $\widetilde{Q}_{k_1^0}(\alpha)$) according as the null H_0, or the alternative H_1, is satisfied. According to the theorem, $\widehat{Q}_{k_1^0}(\alpha)$ drifts to zero under the alternative; conversely, (3.13) entails that, for sufficiently large (p_1, p_2, T),[4] $\widehat{Q}_{k_1^0}(\alpha)$ is bounded away from zero with lower bound

$$\widehat{Q}_{k_1^0}(\alpha) \geq 1 - \alpha - \sqrt{\alpha(1-\alpha)}\sqrt{\frac{2\ln\ln S}{S}}. \tag{3.15}$$

This gap can be exploited to construct a decision rule based on $\widehat{Q}_{k_1^0}(\alpha)$ (or $\widetilde{Q}_{k_1^0}(\alpha)$), not rejecting the null when $\widehat{Q}_{k_1^0}(\alpha)$ (or $\widetilde{Q}_{k_1^0}(\alpha)$) exceeds a threshold, and rejecting otherwise. In theory, one could use the threshold defined in (3.13), but this, albeit valid asymptotically, is likely to be overly conservative in finite samples. A less conservative decision rule in favour of the null could be

$$\widehat{Q}_{k_1^0}(\alpha) \geq (1-\alpha) - f(S), \quad \text{or} \quad \widetilde{Q}_{k_1^0}(\alpha) \geq (1-\alpha) - f(S) \tag{3.16}$$

with $f(S)$ a user-specified, non-increasing function of S such that

$$\lim_{S\to\infty} f(S) = 0 \text{ and } \limsup_{S\to\infty} (f(S))^{-1}\sqrt{\frac{2\ln\ln S}{S}} = 0. \tag{3.17}$$

We call such a family of rules "strong rules", since they originate from a "strong" result (the Law of the Iterated Logarithm). Whilst we discuss possible choices of $f(S)$ in Sections 4 and 5, offering guidelines based on synthetic and real data, here we note that a typical family of default choices for $f(S)$ is

$$f(S) = S^{-q}, \tag{3.18}$$

where $0 < q < 1/2$. On account of (3.16), it can be expected that as q increases, $f(S)$ vanishes more quickly, making the threshold more exacting. As a consequence, tests become less and less conservative, thus leading to a higher probability of rejection of the null hypothesis; in turn, this results in a potential underestimation of k_1^0 in finite samples. Conversely, lower values of q entail that $f(S)$ is larger, whence more conservative tests and, consequently, a higher probability of overstating the number of common factors, at least in finite samples. Based on Monte Carlo evidence, our recommended choice is based on using $q = 1/4$ in (3.18).

Finally, and along similar lines as the comment above, the theorem only requires that M and S be of the same order of magnitude as each other, and that they are (at most) proportional to T. The choice of these tuning parameters is explored in Section 4 (where we show that results are robust to these specifications); here, we note that a default choice is $M = S = T$.

3.3. Determining the number of common factors

The output of the decision rules proposed in (3.16) can be used for two purposes. Firstly, it is possible to check whether $k_1 = 0$ or $k_2 = 0$: this entails that there exists no factor structure in the rows or columns. Similarly, finding $k_1 = k_2 = 0$ implies that there is no factor structure along the row and/or column sections. As a second application of (3.16), upon finding that $k_1 > 0$ (or $k_2 > 0$), the individual decision rules proposed above can be cast in a sequential procedure to determine the number of common row and column factors, based on M_c (and M_r) and $\widetilde{M}_c$ (and $\widetilde{M}_r$) respectively.

As mentioned above, using $\widetilde{M}_c$ and $\widetilde{M}_r$ should yield better results in the presence of a genuine two-way structure. Hence, $\widetilde{M}_c$ and $\widetilde{M}_r$ should be employed if $k_2 > 0$ and $k_1 > 0$ respectively.

The estimator of k_1 (denoted as $\widehat{k}_1$ when using $\widehat{\Psi}_1$ and $\widehat{Q}_{k_1^0}(\alpha)$, and $\widetilde{k}_1$ when using $\widetilde{\Psi}_1$ and $\widetilde{Q}_{k_1^0}(\alpha)$) is the output of the following algorithm:

[4] Formally, (3.13) states that there exists a triple of random variables $(p_{1,0}, p_{2,0}, T_0)$ such that (3.15) holds for all (p_1, p_2, T) with $p_1 \geq p_{1,0}$, $p_2 \geq p_{2,0}$ and $T \geq T_0$.

Step 1 Run the test for $H_0 : k_1 \geq 1$ based on either $\widehat{Q}_1(\alpha)$ or $\widetilde{Q}_1(\alpha)$. If the null is rejected with (3.16), set $\widehat{k}_1 = 0$ (resp. $\widetilde{k}_1 = 0$) and stop, otherwise go to the next step.

Step 2 For $j \geq 2$, run the test for $H_0 : k_1 \geq j$ based on either $\widehat{Q}_j(\alpha)$ or $\widetilde{Q}_j(\alpha)$, constructed using an artificial sample $\left\{\eta_{j,s}^{(m)}, 1 \leq m \leq M\right\}$ generated independently across $1 \leq s \leq S$, and independently of $\left\{\eta_{1,s}^{(m)}, 1 \leq m \leq M\right\}$, ..., $\left\{\eta_{j-1,s}^{(m)}, 1 \leq m \leq M\right\}$. If the null is rejected with (3.16), set $\widehat{k}_1 = j-1$ (resp. $\widetilde{k}_1 = j-1$) and stop; otherwise repeat step 2 until the null is rejected, or until a pre-specified value $k_{\max}$ is reached.

The consistency of $\widehat{k}_1$ and $\widetilde{k}_1$ is stated in the next theorem.

Theorem 4. *We assume that the assumptions of Theorem 3 are satisfied, and that $k_1 \leq k_{\max}$. Then, as $\min\{p_1, p_2, T, M, S\} \to \infty$, it holds that $P^*\left(\widehat{k}_1 = k_1\right) = 1$ for almost all realisations of $\{X_t, 1 \leq t \leq T\}$. If, further, $k_2 > 0$, then it holds that $P^*\left(\widetilde{k}_1 = k_1\right) = 1$, for almost all realisations of $\{X_t, 1 \leq t \leq T\}$.*

3.4. Remarks

We discuss two aspects of our methodology: the estimation of the number of row/column factors when the number of column/row factors is unknown *a priori*, and the performance of our methodology in the presence of weak factors.

3.4.1. Estimation of k_1 when k_2 is unknown

The results above are based on the (implicit) assumption that, when estimating the number of row factors k_1, the number of column factors k_2 is known (and vice versa). In practice, an estimate of k_2 is required prior to computing $\widetilde{k}_1$. This can be obtained using any of the available techniques, but for argument's sake we focus on using $\widehat{k}_2$ derived from using our sequential approach based on $\widehat{\Psi}_{k_2^0}$: whilst suboptimal (as our Monte Carlo shows), $\widehat{k}_2$ is a consistent estimator of k_2 according to Theorem 4; further, its implementation does not require any prior knowledge of k_1, thus being "ready to use" as an initial estimate of k_2.

Further, $\widetilde{M}_c$ and $\widetilde{M}_r$ should be used only if $k_2 > 0$ and $k_1 > 0$ respectively. If this is not known a priori, we recommend using a two-stage approach to estimate k_2 (or, respectively, k_1). In the first step, the applied user should run the test for $H_0 : k_2 \geq 1$, based on $\widehat{\phi}_{k_2^0}$ with $k_2^0 = 1$; according to Theorem 1, this can be applied irrespective of whether $k_1 = 0$ or $k_1 > 0$, and the conclusions from this test are, therefore, robust to the actual value of k_1. Upon rejecting the null, this entails that $k_2^0 = 0$, and therefore a two-way structure does not exists. Consequently, k_1 should be determined using $\widehat{\phi}_{k_1^0}$, based on the spectrum of the "flattened" matrix M_c. If, conversely, the null $H_0 : k_2 \geq 1$ is not rejected, then there is a factor structure across the columns of X_t, and the applied user can use either $\widehat{\phi}_{k_1^0}$ or $\widetilde{\phi}_{k_1^0}$, based on M_c and $\widetilde{M}_c$ respectively. Indeed, when $k_2 > 0$, Theorems 1 and 2 stipulate that both approaches are valid, although (as discussed above) it can be expected that using the spectrum of $\widetilde{M}_c$ should lead to improvements. Indeed, upon finding that $k_1 > 0$, k_2 can also be re-estimated using $\widetilde{\phi}_{k_2^0}$, thus having a (potentially) refined estimator; conversely, if k_1 is found to be zero, k_2 can be estimated using $\widehat{\phi}_{k_2^0}$. As a final remark, in Section B.3.4 in the Supplementary Material we explore the case where $k_2 = 0$, but the applied user employs the projection method anyway, e.g. due to an incorrect initial estimation of k_2, showing that results are anyway robust to this form of mis-specification.

In the next result, we show that, when estimating k_1 using $\widehat{k}_2$, the resulting estimator is consistent, and it preserves the same mode of convergence as in Theorem 4.

Corollary 1. *We assume that the assumptions of Theorem 4 are satisfied and that $\widehat{k}_2$ is used as an estimator of k_2. Then, as $\min\{p_1, p_2, T, M, S\} \to \infty$, it holds that $P^*\left(\widehat{k}_1 = k_1\right) = 1$ for almost all realisations of $\{X_t, 1 \leq t \leq T\}$. If, further, $\widehat{k}_2 > 0$, then it holds that $P^*\left(\widetilde{k}_1 = k_1\right) = 1$, for almost all realisations of $\{X_t, 1 \leq t \leq T\}$.*

3.4.2. Extensions to the case of weak factors

According to Assumption B2, the row and column factors considered in this paper are "pervasive" or "strong". Technically, this is due to the fact that the (squared) L_2-norm of the loading matrices diverge at rates $\|R'R\| = \Omega(p_1)$ and $C'C = \Omega(p_2)$. However, the literature on vector factor models has recently considered the case of common factors where the square of the L_2-norm of the loading matrix still diverges, but at a rate slower than the cross-sectional dimension — in our context, this would e.g. correspond to having $\|R'R\| = \Omega\left(p_1^{\alpha_1}\right)$, for some $0 < \alpha_1 < 1$. The recent contributions by Uematsu and Yamagata (2021) and Uematsu and Yamagata (2022), and the references cited therein, offer a state-of-the-art discussion of the issue of determining weak factors in the context of vector factor models. As far as matrix factor models are concerned, the literature has also considered the possible presence of weak factors: for example, Wang et al. (2019), Chen et al. (2020a) and Gao and Tsay (2021) all consider weak factors along both the row and column spaces, modelling their strength with an approach similar to the one in this paper; and, as discussed in the introduction, Lam (2021) allows for the presence of weak common factors in the idiosyncratic error term, thereby allowing for factor-induced cross-correlation.

Hence, in this section we briefly investigate how our methodology works in the presence of weak factors. For the sake of a concise discussion, we focus primarily on determining the presence of only one weak common factor ($k_1 = 1$) in the case where both the row and column common factors are possibly weak.

In the context of factor models for matrix valued data, we assume that Assumptions B1–C1 above all hold, but we replace/integrate Assumption B2 with the following

Assumption B6. Assumption B2 holds with part (ii) replaced by $\|p_1^{-\alpha_1}R'R - I_{k_1}\| \to 0$ as $p_1 \to \infty$, for some $0 < \alpha_1 \leq 1$, and $\|p_2^{-\alpha_2}C'C - I_{k_2}\| \to 0$ as $p_2 \to \infty$, for some $0 < \alpha_2 \leq 1$.

This assumption summarises the discussion at the beginning of this section: the (squared) L_2-norm of the loading matrix R diverges, at a rate that is possibly *lower* than p_1. Similarly, the column factors can also be weak. Prior to reporting the main theoretical result, we offer a heuristic preview of the main arguments. Repeating the proofs of our main results, it can be shown that, for all $j \leq k_1$

$$\widehat{\lambda}_j = \Omega_{a.s.}\left(p_1^{\alpha_1}p_2^{\alpha_2-1}\right) \quad \text{and} \quad \widetilde{\lambda}_j = \Omega_{a.s.}\left(p_1^{\alpha_1}p_2^{\alpha_2-1}\right),$$

whereas the conclusions of Theorems 1 and 2 still hold true for all $j > k_1$. All methodologies that try to determine the number of common factors require some eigen-gap in the second order matrices. This entails that – in the case of our methodology – detection of weak factors is in principle possible as long as, as $\min\{p_1, p_2, T\} \to \infty$

$$\frac{\widehat{\lambda}_{k_1}}{\widehat{\lambda}_{k_1+1}} \xrightarrow{a.s.} \infty. \tag{3.19}$$

The following result summarises the ability of $\widehat{k}_1$ to estimate the number of common factors in the presence of weak factors.

Corollary 2. *We assume that the assumptions of Theorem 3 hold, with Assumption B2 replaced with Assumption B6. Then, as* $\min\{p_1, p_2, T, M, S\} \to \infty$, *if*

$$p_1^{\alpha_1}p_2^{\alpha_2-1} \to \infty, \tag{3.20}$$

$$p_1^{\alpha_1-1}p_2^{\alpha_2-1/2}T^{1/2} \to \infty, \tag{3.21}$$

then it holds that $P^*\left(\widehat{k}_1 = k_1\right) = 1$, *for almost all realisations of* $\{X_t, 1 \leq t \leq T\}$.

The results in Corollary 2 can be read in conjunction with similar results in Wang et al. (2019), Chen et al. (2020a) and Gao and Tsay (2021). Some comments on (3.20) and (3.21) are in order. A quick inspection of Theorem 1 reveals that the leading eigenvalues of M_c are proportional to $\Omega\left(p_1^{\alpha_1}p_2^{\alpha_2-1}\right)$; hence, (3.20) ensures that such eigenvalues diverge, which is a necessary condition to find factors. Eq. (3.21) entails that (3.19) holds, and therefore the "signal" associated to common factors is not drown out by the estimation "noise". In this respect, (3.21) states, heuristically, that when p_1 is "too big" (in comparison with the other dimensions, p_2 and T), detection of weak factors is less easy.

Eq. (3.21) can be illustrated through some examples. Consider, for simplicity, $\alpha_2 = 1$ – that is, column factors are strong. If $p_1 = p_2 = T$, then (3.21) boils down to $p_1^{\alpha_1} \to \infty$, which is always true as long as $\alpha_1 > 0$: this entails that, in this case, arbitrarily weak factors can be (potentially) detected. Indeed, in (3.4), $\widehat{\lambda}_{k_1^0}$ is dampened by a factor $p_1^{-\delta}$, with – in this example – $\delta = \varepsilon$. Hence, it can be verified that $\widehat{\phi}_{k_1^0}$ will diverge as long as $\alpha_1 > \varepsilon$: the choice of ε, which is entirely up to the applied user, will determine which weak factors can be detected and which ones will be left out. As another example, consider a "very long" matrix sequence, where e.g. $p_1 = T^{1/2}$; in such a case, (3.21) becomes $p_1^{\alpha_1}p_2^{1/2} \to \infty$, which is satisfied even when p_2 diverges very slowly (e.g., even if $p_2 = \ln p_1$); in this case, again, the choice of ε will determine which weak factors can be detected, and this is further enhanced the larger T is. Conversely, consider the case where the matrix sequence is "very short", e.g. $T = p_1^{1/2}$. In such a case, (3.21) becomes $p_1^{\alpha_1-3/4}p_2^{1/2} \to \infty$, which, if e.g. $p_2 = p_1$, entails that detection is possible only when $\alpha_1 > 1/4$: very weak factors cannot be detected in this case. A similar phenomenon was also noted in the case of vector valued series by Trapani (2018): however, in the case of matrix valued series, a small T can be offset by a large value of p_2, which is a major advantage of having a matrix structure in the data. Other examples can also be considered, but the general message is that detection of weak factors is helped by both T and p_2.

In the presence of common factors that are weak along the columns, i.e. when $\alpha_2 < 1$, the interpretation of (3.21) is more convoluted, but essentially the same. As mentioned above, large values of p_2 help the estimation of k_1: however, such helpfulness is dampened when the common factors in the column are weak – in essence, because the information coming from aggregating the columns is, itself, weak.

As a final remark, we note that the case of using $\widetilde{\lambda}_{k_1}$ is more complicated, essentially because the estimation error of $\widehat{C}$ is compounded (and inflated) by the presence of weak factors. In the interest of brevity, we relegate the treatment of this case to Lemmas C.4 and C.5 in the Supplementary Material. The latter result is, essentially, an equivalent of Theorem 2 in the presence of weak factors. An analogue restriction to (3.21) can be derived from Lemma C.5; in particular,

Eq. (C.14) in the lemma suggests that detection of weak factors requires the necessary condition $p_1^{\alpha_1} p_2^{\alpha_2 - 1} \to \infty$. Sufficient conditions, similar to (3.21), can be derived from equation (C.14). However, results are far more complicated, and of dubious helpfulness. Technically, this is due to the fact that weak factors also enter the projected estimator $\widehat{C}$, making it less precise (again, due to the fact that estimation of C is now based on a weaker "signal").

4. Simulation studies

In this section, we evaluate the finite sample performances of our strong rule to determine whether there is a two-way or one-way factor structure, and of the sequential procedure to estimate the number of common factors. As far as the latter is concerned, we compare our Sequential Testing Procedures (henceforth denoted as "STP") with several competing methodologies available in the literature.

We begin with describing the implementation of our procedures. For the proposed STP, three different approaches can be adopted: firstly, the test statistics can be constructed using the eigenvalues of $M_{\mathcal{L}}$ (or $M_{\mathcal{L}}$), and we denote this approach as STP_1; secondly, the test statistics can be constructed using the eigenvalues of $\widetilde{M}_c$ (or $\widetilde{M}_r$), using STP_1 as a preliminary step to estimate e.g. k_2 and subsequently using the estimated value, $\widehat{k}_2$, to construct the initial estimator $\widehat{C}$, and we denote this approach as STP_3; and, finally, the test statistics can still be constructed using the eigenvalues of $\widetilde{M}_c$ (or $\widetilde{M}_r$), but in order to avoid the (finite sample) risk of understating k_2 one can use a deliberately large number instead of the STP_1 estimator (we set $\widehat{k}_2 = k_{\max} = 8$), and we denote this approach as STP_2.

When computing integrals such as $\int_{-\infty}^{\infty} \left[\widehat{v}_{k_1^0}(u)\right]^2 dF(u)$, we use the distribution of the standard normal as weight function $F(u)$, using a Gauss–Hermite quadrature with

$$\widehat{\Psi}_{k_1^0} = \sum_{s=1}^{n_S} w_s \widehat{v}_{k_1^0}(\sqrt{2} z_s). \tag{4.1}$$

In (4.1), the z_ss, $1 \le s \le n_S$, are the zeros (of the physicist's version) of the Hermite polynomial $H_{n_S}(z)$ defined as

$$H_{n_S}(z) = (-1)^{n_S} \exp\left(x^2\right) \frac{d^{n_S}}{dx^{n_S}} \exp\left(-x^2\right), \tag{4.2}$$

and the weights w_s are defined as

$$w_s = \frac{2^{n_S - 1} (n_S - 1)!}{n_S \left[H_{n_S - 1}(z_s)\right]^2}.$$

Thus, when computing $\widehat{v}_{k_1^0}(u)$ in Step 2 of the algorithm, we construct n_S of these statistics, each using $u = \pm\sqrt{2} z_s$. The values of the roots z_s, and of the corresponding weights w_s, are tabulated e.g. in Salzer et al. (1952). In our case, we have used $n_S = 4$, which corresponds to $w_1 = w_4 = 0.05$ and $w_2 = w_3 = 0.45$, and $u_1 = -u_4 = 2.4$ and $u_2 = -u_3 = 0.7$.

Data generation

We use the same Data Generating Process (DGP) as Yu et al. (2022) in order to generate X_t. Specifically, when $k_1 \neq 0$ and $k_2 \neq 0$, i.e., when there is a factor structure, we generate the entries of R and C independently from the uniform distribution $\mathcal{U}(-1, 1)$, and we let

$$\text{Vec}(F_t) = \phi \text{Vec}(F_{t-1}) + \sqrt{\theta\left(1 - \phi^2\right)} \text{Vec}(\epsilon_t), \quad \epsilon_t \sim i.i.d.\ \mathcal{N}\left(0, I_{k_1 \times k_2}\right), \tag{4.3}$$

$$\text{Vec}(E_t) = \psi \text{Vec}(E_{t-1}) + \sqrt{1 - \psi^2} \text{Vec}(U_t), \quad U_t \sim i.i.d.\ \mathcal{N}(0, V_E \otimes U_E), \tag{4.4}$$

where U_E and V_E are matrices with ones on the diagonal, and the off-diagonal entries are a/p_1 and a/p_2, respectively. The parameter a controls cross-sectional dependence, with larger a leading to stronger cross-dependence; we have used $a = 2$ in our simulations. In the case that no factor structure exists, we simply let $X_t = E_t$, with E_t generated in the same way as in (4.3)–(4.4). The parameters ϕ and ψ control both the temporal and cross-sectional correlations of X_t; with nonzero ϕ and ψ, the generated factors are temporally correlated while the idiosyncratic noises are both temporally and cross-sectionally correlated. In all our experiments, we set $\phi = \psi = 0.1$ and use $\theta = 1$.[5] In all the simulation settings, the reported results are based on 500 replications.

As a final remark, our DGP entails that, letting $\Sigma_E = E\left(\text{Vec}(E_t)\text{Vec}'(E_t)\right)$, and $\Sigma_C = E\left(\text{Vec}\left(RF_tC'\right)\text{Vec}'\left(RF_tC'\right)\right)$, the signal-to-noise ratio is given by

$$\sigma = \frac{\sum_{i,j=1}^{p_1 p_2} \{\Sigma_C\}_{ij}}{\sum_{i,j=1}^{p_1 p_2} \{\Sigma_E\}_{ij}} \simeq \theta \frac{k_1 k_2}{9(1 + a)^2}, \tag{4.5}$$

for large values of p_1 and p_2. We point out that, in various unreported simulations (and in a limited set of further Monte Carlo exercises reported in Section 4.5 in the Supplement) we have tried to alter θ; as expected, when this increases

[5] We have also tried $\phi = \psi = 0.3$, and results are essentially the same.

Table 1
Proportions of correctly determining whether there exists factor structure using $\widehat{Q}_1(\alpha)$ and $\widetilde{Q}_1(\alpha)$ over 500 replications with $M = S = 300$, $\phi = \psi = 0.1$, $f(S) = S^{-1/4}$.

α	(k_1, k_2)	Method	$(p_1, T) = (100, 100)$			$(p_1, T) = (150, 150)$		
			$p_2 = 15$	$p_2 = 20$	$p_2 = 30$	$p_2 = 15$	$p_2 = 20$	$p_2 = 30$
0.01	(0, 0)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	1	1	1	1	1	1
	(1, 1)	$\widehat{Q}_1(\alpha)$	0.648	0.838	0.97	0.850	0.946	0.998
		$\widetilde{Q}_1(\alpha)$	0.970	0.998	1.00	0.996	1.000	1.000
		Vec	0.000	0.000	0.00	0.000	0.000	0.000
	(1, 3)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	0	0	0	0	0	0
0.05	(0, 0)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	1	1	1	1	1	1
	(1, 1)	$\widehat{Q}_1(\alpha)$	0.334	0.574	0.904	0.644	0.84	0.984
		$\widetilde{Q}_1(\alpha)$	0.886	0.992	1.000	0.982	1.00	1.000
		Vec	0.000	0.000	0.000	0.000	0.00	0.000
	(1, 3)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	0	0	0	0	0	0
0.10	(0, 0)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	1	1	1	1	1	1
	(1, 1)	$\widehat{Q}_1(\alpha)$	0.210	0.422	0.786	0.52	0.724	0.966
		$\widetilde{Q}_1(\alpha)$	0.812	0.992	1.000	0.96	1.000	1.000
		Vec	0.000	0.000	0.000	0.00	0.000	0.000
	(1, 3)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	0	0	0	0	0	0

all criteria improve (this is particularly evident for the estimator proposed by Lam, 2021), but the relative performance remains unaltered.

Determining whether there is a factor structure

We investigate the finite sample performance of our "strong" rule to determine whether a factor structure exists in the matrix time-series data. This offers a solution to the question in the discussion section of Yu et al. (2022): *is the matrix factor structure true for the time series?*

We study two scenarios: first, the case of "small" p_2, using $p_1 = T = \{100, 150\}$ and $p_2 = \{15, 20, 30\}$; second, the more balanced cases $(p_1, p_2, T) = (50, 50, 50)$, $(100, 100, 50)$, $(150, 150, 50)$, $(50, 50, 100)$, $(100, 100, 100)$, and $(150, 150, 100)$.[6] In all our simulations, we use $k_{\max} = 8$, although we tried different values for $k_{\max}$ and the results show that the proposed methods are not sensitive to the choice of it. As far as our decision rule is concerned, we have used $\alpha \in \{0.01, 0.05, 0.10\}$, $M = 300$ and $S = 300$; in (3.16), we have used $f(S) = S^{-1/4}$. Results using different combinations of (α, M, S) and different choice of $f(S)$ are in the Supplementary Material. By way of comparison, we also use the test developed in Trapani (2018), applying it to the $p_1 p_2 \times 1$ series $\text{Vec}(X_t)$ — this is denoted by "Vec" in our tables.

Firstly, we consider the case of the existence of a matrix factor structure by setting $(k_1, k_2) = \{(1, 1), (1, 3)\}$ in (4.3). We report the proportions of correctly claiming that there exists factor structure by the "strong" rule in the second and third rows of each subpanel of Tables 1 and 2; $\widetilde{Q}_1$ (using $\widetilde{\Psi}_1^S$) and $\widehat{Q}_1$ (using $\widehat{\Psi}_1^S$) denote our procedure with and without projection technique, respectively. The results indicate that, when using $\widetilde{\Psi}_1^S$, a factor structure is found more than 95% of the times whenever $p_1 \geq 100$, with the sole exception of the (small sample) case $(p_1, p_2, T) = (100, 15, 100)$. Results are always worse when $\widehat{\Psi}_1^S$ is used, although improvements are seen for larger sample sizes (see Table 2): this reinforces the case in favour of the projection method developed by Yu et al. (2022), especially when p_2 is small. When $p_1 = 50$, results are essentially the same, with few exceptions. All across the board, results obtained using the test by Trapani (2018) are very bad, indicating that, in essence, the test always fails to detect the existence of a factor structure when this is present. This result is not entirely unexpected: based on Trapani (2018), eigenvalues are scaled by a factor $(p_1 p_2)^{-\delta}$, where $\delta = 0.685$ in the "best" case where $p_1 = 100$ and $p_2 = 15$. Indeed, δ increases (by construction) with p_2, thus resulting in dampening eigenvalues even more. In turn, this makes detection of diverging eigenvalues particularly difficult.

[6] In Section B.3 in the supplement, we complement these results with two more sets of experiments, considering p_1 and p_2 being comparable and considering p_1 and p_2 being comparable and small, respectively. Results are broadly similar to the ones reported here.

Table 2
Proportions of correctly determining whether there exists factor structure using $\widehat{Q}_1(\alpha)$ and $\widetilde{Q}_1(\alpha)$ over 500 replications with $M = S = 300$, $\phi = \psi = 0.1$, $f(S) = S^{-1/4}$.

α	(k_1, k_2)	Method	$T = 50$			$T = 100$		
			$p_1 = p_2 = 50$	$p_1 = p_2 = 100$	$p_1 = p_2 = 150$	$p_1 = p_2 = 50$	$p_1 = p_2 = 100$	$p_1 = p_2 = 150$
0.01	(0, 0)	$\widehat{Q}_1(\alpha)$	1.000	1	1	1.000	1.000	1.000
		$\widetilde{Q}_1(\alpha)$	0.142	1	1	0.766	0.166	0.988
		Vec	1.000	1	1	1.000	1.000	1.000
	(1, 1)	$\widehat{Q}_1(\alpha)$	0.986	1	1	0.99	1	1
		$\widetilde{Q}_1(\alpha)$	1.000	1	1	1.00	1	1
		Vec	0.000	0	0	0.00	0	0
	(1, 3)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	0	0	0	0	0	0
0.05	(0, 0)	$\widehat{Q}_1(\alpha)$	1.000	1	1	1	1.000	1
		$\widetilde{Q}_1(\alpha)$	0.938	1	1	1	0.996	1
		Vec	1.000	1	1	1	1.000	1
	(1, 1)	$\widehat{Q}_1(\alpha)$	0.888	0.998	1	0.936	1	1
		$\widetilde{Q}_1(\alpha)$	1.000	1.000	1	1.000	1	1
		Vec	0.000	0.000	0	0.000	0	0
	(1, 3)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	0	0	0	0	0	0
0.10	(0, 0)	$\widehat{Q}_1(\alpha)$	1.000	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	0.998	1	1	1	1	1
		Vec	1.000	1	1	1	1	1
	(1, 1)	$\widehat{Q}_1(\alpha)$	0.794	0.972	1	0.846	1	1
		$\widetilde{Q}_1(\alpha)$	1.000	1.000	1	1.000	1	1
		Vec	0.000	0.000	0	0.000	0	0
	(1, 3)	$\widehat{Q}_1(\alpha)$	1	1	1	1	1	1
		$\widetilde{Q}_1(\alpha)$	1	1	1	1	1	1
		Vec	0	0	0	0	0	0

Secondly, we investigate the performance of our proposed methodology when there is no factor structure in the matrix time series $X_t = E_t$, i.e. $(k_1, k_2) = (0, 0)$, with E_t generated in the same way as in (4.3). In the first row of each subpanel of Tables 1 and 2, we report the proportions of correctly claiming that there exists no factor structure, from which we can conclude that our proposed methodology is extremely powerful in identifying the absence of a factor structure, even in the small sample case $(p_1, p_2, T) = (100, 15, 100)$, and with or without projection. There are some puzzling exceptions in Table 2, especially when using $\alpha = 0.01$, but these issues vanish as the sample size increases. We note again that results obtained using the test by Trapani (2018) are, in this case, very satisfactory, but this is clearly a spurious effect due to the reasons discussed above.

In the Supplementary Material, we report more results obtained under different scenarios, which reinforce our conclusions — see e.g. Table B.5, where we consider smaller values of p_1 and p_2. In Tables B.10 and B.11, we report results based on different combinations of (α, M, S). Results are essentially the same when using $\widetilde{\Psi}_1^S$, whereas $\widehat{\Psi}_1^S$ is more sensitive (at least in small samples) to the choice of M and – albeit to a lesser extent – S. In particular, as far as the former is concerned, smaller values of it seem to yield better results in small samples. Finally, in Table B.12 in the Supplementary Material, we assess the sensitivity to $f(S)$; as can be expected, results are affected by the choice of the threshold, but this is only marginal when using $\widetilde{\Psi}_1^S$ and, again, more pronounced when using $\widehat{\Psi}_1^S$.

Determining the number of common factors

We investigate the finite sample performances of the sequential testing procedure introduced in Section 3.3. We use the same design, and consider the same combinations of (p_1, p_2, T) as in the previous set of experiments. In order to evaluate the sequential procedure, we use it to estimate the number of row factors k_1 considering the cases $(k_1, k_2) = \{(1, 1), (1, 3), (3, 1), (3, 3)\}$. As in the previous section, we use $M = 300$ and $S = 300$, and $f(S) = S^{-1/4}$, and we only report results for the case $\alpha = 0.01$ for brevity.

As well as assessing the performance of the STP methods, we compare these against the most popular alternatives in the literature. We have considered the following techniques:[7] the Iterative Eigenvalue-Ratio (denoted as IterER) studied in Yu et al. (2022); the α-PCA Eigenvalue-Ratio method (denote as α-PCA) proposed by Chen and Fan (2021); the iterative versions of the Eigenvalue Ratio and of the Information Criteria algorithms by Han et al. (2022) (here denoted as iTIP-ER and iTIP-IC respectively); and the method proposed by Lam (2021) (denoted as TCorTh). Finally, we have used the

[7] Details on how each procedure has been implemented are in Section B.1 in the Supplement.

Table 3
Simulation results for estimating k_1 in the form $x(y|z)$, x is the sample mean of the estimated factor numbers based on 500 replications $\alpha = 0.01, M = S = 300, \phi = \psi = 0.1, f(S) = S^{-1/4}$, y and z are the proportions of under- and exact-determination of the factor number, respectively.

(k_1, k_2)	Method	$(p_1, T) = (100, 100)$			$(p_1, T) = (150, 150)$		
		$p_2 = 15$	$p_2 = 20$	$p_2 = 30$	$p_2 = 15$	$p_2 = 20$	$p_2 = 30$
(1, 1)	STP_1	0.648(0.352|0.648)	0.838(0.162|0.838)	0.97(0.03|0.97)	0.85(0.15|0.85)	0.946(0.054|0.946)	0.998(0.002|0.998)
	STP_2	0.968(0.032|0.968)	0.998(0.002|0.998)	1(0|1)	0.998(0.002|0.998)	1(0|1)	1(0|1)
	STP_3	0.706(0.294|0.706)	0.914(0.086|0.914)	0.994(0.008|0.99)	0.758(0.242|0.758)	0.942(0.058|0.942)	0.994(0.006|0.994)
	IterER	0.474(0.526|0.474)	0.766(0.234|0.766)	0.972(0.028|0.972)	0.544(0.456|0.544)	0.82(0.18|0.82)	0.988(0.012|0.988)
	α-PCA	0.796(0.204|0.796)	0.832(0.168|0.832)	0.884(0.116|0.884)	0.956(0.044|0.956)	0.962(0.038|0.962)	0.986(0.014|0.986)
	iTIP-IC	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	iTIP-ER	1.368(0|0.636)	1.35(0|0.65)	1.336(0|0.664)	1.39(0|0.61)	1.328(0|0.672)	1.352(0|0.648)
	TCorTh	2(0|0)	2(0|0)	2(0|0)	2(0|0)	2(0|0)	2(0|0)
	IC	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
(1, 3)	STP_1	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	STP_2	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	STP_3	0.988(0.012|0.988)	1(0|1)	1(0|1)	0.996(0.004|0.996)	1(0|1)	1(0|1)
	IterER	0.97(0.03|0.97)	0.996(0.004|0.996)	1(0|1)	0.996(0.004|0.996)	1(0|1)	1(0|1)
	α-PCA	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	iTIP-IC	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	iTIP-ER	1.186(0|0.814)	1.174(0|0.826)	1.144(0|0.856)	1.12(0|0.88)	1.098(0|0.902)	1.098(0|0.902)
	TCorTh	2(0|0)	2(0|0)	2(0|0)	2(0|0)	2(0|0)	2(0|0)
	IC	2.858(0.142|0.858)	2.954(0.046|0.954)	2.988(0.012|0.988)	2.968(0.032|0.968)	2.994(0.006|0.994)	3(0|1)
(3, 1)	STP_1	2.106(0.338|0.662)	2.732(0.12|0.88)	2.994(0.002|0.998)	2.71(0.106|0.894)	2.946(0.018|0.982)	3(0|1)
	STP_2	2.944(0.024|0.976)	3(0|1)	3(0|1)	2.982(0.006|0.994)	3(0|1)	3(0|1)
	STP_3	3(0|1)	3(0|1)	3(0|1)	2.994(0.002|0.998)	3(0|1)	3(0|1)
	IterER	2.988(0.004|0.996)	3(0|1)	3(0|1)	2.982(0.006|0.994)	3(0|1)	3(0|1)
	α-PCA	2.404(0.2|0.8)	2.494(0.17|0.83)	2.766(0.084|0.902)	2.862(0.046|0.954)	2.94(0.02|0.98)	2.934(0.022|0.978)
	iTIP-IC	1(1|0)	1(1|0)	1(1|0)	1(1|0)	1(1|0)	1(1|0)
	iTIP-ER	2.646(0.446|0.346)	2.692(0.448|0.312)	2.642(0.436|0.376)	2.736(0.45|0.328)	2.924(0.376|0.364)	2.782(0.37|0.424)
	TCorTh	3.986(0|0.014)	3.988(0|0.012)	3.988(0|0.012)	3.988(0|0.012)	4(0|0)	3.998(0|0.002)
	IC	2.994(0.004|0.996)	2.994(0.006|0.994)	3(0|1)	2.996(0.002|0.998)	3(0|1)	3(0|1)
(3, 3)	STP_1	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	STP_2	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	STP_3	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	IterER	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	α-PCA	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	iTIP-IC	1(1|0)	1(1|0)	1(1|0)	1(1|0)	1(1|0)	1(1|0)
	iTIP-ER	2.534(0.426|0.478)	2.69(0.366|0.48)	2.71(0.324|0.57)	2.718(0.304|0.606)	2.824(0.23|0.686)	2.842(0.218|0.694)
	TCorTh	3.004(0|0.996)	3.002(0|0.998)	3.002(0|0.998)	3.012(0|0.988)	3(0|1)	3(0|1)
	IC	8.458(0.33|0.67)	8.69(0.234|0.766)	8.908(0.076|0.924)	8.89(0.064|0.936)	8.972(0.022|0.978)	8.996(0.004|0.996)

Note that IC determines the total number of factors, i.e. $k_1 k_2$.

information criterion $PC_{p1}(k)$ proposed in Bai and Ng (2002), applied to a vectorised version of X_t, to determine the total number of factors $k = k_1 k_2$ (this is denoted by IC).

Results are in Tables 3 and 4, from which we can draw the following three conclusions.

First, especially for the case $k_1 = k_2 = 1$, and especially when p_2 is small, the STP_2 and STP_3 procedures dominate the IterER and the α-PCA procedures, which have a pronounced tendency to understate the number of common factors — thus, in this case, mistakenly finding no evidence of a row factor structure and, consequently, mistakenly indicating a vector, as opposed to a matrix, factor model. This is not true for the iTIP-IC and the iTIP-ER procedures, which always correctly estimate k_1 as equal to 1; however, these procedures (especially iTIP-IC) are less able to determine the presence of further common factors when $k_1 = 3$. Note further that, by construction, they are initialised at $k_1 = 1$, so they cannot understate k_1 but are unable to check whether $k_1 = 0$ (although, in principle, it is possible to extend this method to check if there are $k_1 = 0$ factors by constructing an artificial eigenvalue as in Ahn and Horenstein, 2013). In general, the STP_2 procedure (and, to a lesser extent, the STP_3 one) dominates over all other procedures in almost all cases considered, which makes a very strong case to consider the preliminary step of projecting the data X_t onto C prior to undertaking any analysis. We also note that the STP_1 method performs comparably with the α-PCA method, but it is inferior to the IterER method, albeit with some exceptions — e.g. when $k_1 = k_2 = 1$, and $p_2 = 15$. Results, as Table 4 demonstrates, improve as the sample sizes increase, and become comparable with those obtained with other criteria. The results in Table 4 also contain the case of smaller $p_1 = 50$: in such a case, it is evident that reducing p_1 worsens the overall ability of our procedures, which can be explained by noting that a lower p_1 corresponds to the spiked eigenvalues diverging at a slower rate.

Second, all procedures seem to improve as k_2 increases, as can be anticipated in the light of (4.5). In such cases, the STP_2 and STP_3 procedures still retain their advantage especially for small values of p_2, but less evidently than in the previous cases.

Third, confirming what is also found in the previous section, the STP_2 and STP_3 methodologies always outperform the STP_1 one. Indeed, the STP_1 procedure works well for large sample sizes, but it is dominated even by the IterER methodology in small samples (with few exceptions), and, to a lesser extent, by the α-PCA one. This suggests that the gains observed with STP_2 arise from two equally important sources: the use of the projection method proposed by Yu et al. (2022), and the use of our randomised tests *cum* the decision rule advocated in (3.16).

Finally, we note that we have run further experiments in the Supplementary Material to assess the robustness of our procedures. In Table B.6, we consider the case of $p_1 = p_2$ and "small", obtaining, broadly, the same results as above; the

Table 4
Simulation results for estimating k_1 in the form $x(y|z)$, x is the sample mean of the estimated factor numbers based on 500 replications $\alpha = 0.01, M = S = 300, \phi = \psi = 0.1, f(S) = S^{-1/4}$, are the proportions of under- and exact-determination of the factor number, respectively.

(k_1, k_2)	Method	$T = 50$			$T = 100$		
		$p_1 = p_2 = 50$	$p_1 = p_2 = 100$	$p_1 = p_2 = 150$	$p_1 = p_2 = 50$	$p_1 = p_2 = 100$	$p_1 = p_2 = 150$
(1, 1)	STP_1	1.966(0.014|0.006)	1(0|1)	1(0|1)	1.976(0.01|0.004)	2(0|0)	1.05(0|0.95)
	STP_2	1.91(0|0.09)	1.002(0|0.998)	1(0|1)	1.694(0|0.306)	1.082(0|0.918)	1(0|1)
	STP_3	1.028(0|0.972)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	IterER	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	α-PCA	0.558(0.53|0.382)	0.88(0.12|0.88)	0.992(0.008|0.992)	1.144(0.3|0.256)	0.952(0.048|0.952)	1(0|1)
	iTIP-IC	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	iTIP-ER	1.258(0|0.742)	1.31(0|0.69)	40.352(0|0)	1.266(0|0.74)	1.236(0|0.764)	1.236(0|0.764)
	TCorTh	2(0|0)	2(0|0)	1.55(0|0.45)	2(0|0)	2(0|0)	2(0|0)
	IC	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
(1, 3)	STP_1	1.344(0|0.656)	1(0|1)	1(0|1)	1.294(0|0.706)	1.328(0|0.672)	1(0|1)
	STP_2	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	STP_3	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	IterER	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	α-PCA	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	iTIP-IC	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)	1(0|1)
	iTIP-ER	1.192(0|0.808)	1.234(0|0.766)	5.6(0|0.724)	1.084(0|0.916)	1.076(0|0.924)	1.104(0|0.896)
	TCorTh	2(0|0)	1.436(0|0.564)	1(0|1)	2(0|0)	2(0|0)	2(0|0)
	IC	2.78(0.208|0.792)	2.908(0.09|0.91)	2.904(0.094|0.906)	3(0|1)	3(0|1)	3(0|1)
(3, 1)	STP_1	3.13(0.002|0.866)	3(0|1)	3(0|1)	3.086(0.002|0.91)	3.182(0|0.818)	3(0|1)
	STP_2	3.002(0|0.998)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	STP_3	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	IterER	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	α-PCA	3.38(0.142|0.072)	3.054(0.028|0.858)	3(0|1)	3.768(0.044|0.056)	3.084(0.004|0.9)	3(0|1)
	iTIP-IC	1(1|0)	1(1|0)	1(1|0)	1(1|0)	1(1|0)	1(1|0)
	iTIP-ER	2.556(0.434|0.42)	2.778(0.34|0.492)	44.404(0.02|0.058)	2.572(0.45|0.386)	2.818(0.316|0.496)	2.87(0.268|0.57)
	TCorTh	3.844(0|0.156)	3.052(0|0.948)	3(0|1)	3.992(0|0.008)	3.998(0|0.002)	3.784(0|0.216)
	IC	2.776(0.21|0.79)	2.866(0.132|0.868)	2.92(0.078|0.922)	3(0|1)	3(0|1)	3(0|1)
(3, 3)	STP_1	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	STP_2	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	STP_3	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	IterER	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)	3(0|1)
	α-PCA	2.998(0.002|0.998)	3(0|1)	3(0|1)	2.998(0.002|0.998)	3(0|1)	3(0|1)
	iTIP-IC	1(1|0)	1(1|0)	1(1|0)	1(1|0)	1(1|0)	1(1|0)
	iTIP-ER	2.61(0.408|0.456)	2.806(0.282|0.578)	5.514(0.254|0.54)	2.728(0.284|0.632)	2.824(0.228|0.676)	2.906(0.166|0.744)
	TCorTh	3.002(0|0.998)	3(0|1)	3(0|1)	3.102(0|0.898)	3(0|1)	3(0|1)
	IC	7.192(0.888|0.112)	7.832(0.712|0.288)	7.896(0.74|0.26)	8.958(0.038|0.962)	8.998(0.002|0.998)	9(0|1)

Note that IC determines the total number of factors, i.e. $k_1 k_2$.

case of weak factors is studied in Tables B.7 and B.8; finally, in Table B.9, we analyse the case in which $k_2 = 0$, but the applied user mistakenly uses $\widetilde{k}_1$. We also investigate the robustness of our procedure to its specifications. In Table B.3, we assess the robustness to different specifications to the weight function $F(u)$ required in Step 4 of our randomisation algorithm, and in Table B.13, we assess the impact of the threshold function $f(S)$ for the case $(k_1, k_2) = (3, 3)$. The broad conclusion, even in this case, is that our procedure is not affected by these specifications, reinforcing the message that although some specifications need to be chosen by the researcher, the impact thereof is negligible. Finally, in Tables B.15–B.17 we report the computational times of our procedures, comparing them against those of the other criteria considered in the above.

5. Empirical studies

We illustrate our procedure through two applications: we firstly consider a 2-D image recognition dataset (Section 5.1), and then present an application to a set of macroeconomic indices (Section 5.2).

5.1. MNIST: handwritten digit numbers

In our first example, we apply matrix time series to an image recognition dataset, namely the Modified National Institute of Standards and Technology (MNIST) dataset, which has been analysed in numerous applications of classification algorithms and machine learning, and which consists of images of handwritten digit numbers from 0 to 9. As is typical in these applications, each single (grey-scale) image represents the matrix X_t, whose elements are the pixels of the image. We only use the training set, which contains $T = 10{,}000$ images; in our dataset, the digits have been size-normalised and centred in a fixed-size image with 28×28 pixels, thus having $p_1 = p_2 = 28$. We standardise the pixels at each location.

The estimated numbers of row and column factors are reported in Table 5 with multiple combinations of α, S and $f(S)$, as in the previous section. Since p_1 and p_2 are larger than those in our previous example, we use $M = 200$ in the testing (we tried $M = 100$ and 300 and results are essentially the same).

Results and conclusions are similar to those in the previous section. In particular, a bigger difference emerges in the performance of projection versus non-projection based estimation, with the former offering a performance which is more robust across the different specifications. In light also of the results in Section 4, the findings in this section strengthen the case in favour of the projection-based estimator. We note that, when using this technique, the number of row factors is

Table 5
Estimated numbers of row and column factors for handwritten digit number dataset. Five ways to select the threshold. $f_1(S)$: $1-\alpha-\sqrt{2\ln\ln S/S}$; $f_2(S)$: $1-\alpha-S^{-1/3}$; $f_3(S)$: $1-\alpha-S^{-1/4}$; $f_4(S)$: $1-\alpha-S^{-1/5}$; $f_5(S)$: $(1-\alpha)/2$.

S	α	No-projection					Projection				
		$f_1(S)$	$f_2(S)$	$f_3(S)$	$f_4(S)$	$f_5(S)$	$f_1(S)$	$f_2(S)$	$f_3(S)$	$f_4(S)$	$f_5(S)$
200	0.01	(0, 3)	(0, 3)	(0, 3)	(0, 4)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)
300	0.01	(0, 3)	(0, 3)	(0, 3)	(0, 4)	(4, 5)	(4, 5)	(4, 4)	(4, 5)	(4, 5)	(4, 5)
400	0.01	(0, 3)	(0, 3)	(0, 3)	(0, 3)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)
200	0.05	(0, 0)	(0, 0)	(0, 0)	(0, 1)	(0, 1)	(4, 3)	(4, 3)	(4, 3)	(4, 3)	(4, 3)
300	0.05	(0, 0)	(0, 0)	(0, 0)	(0, 1)	(0, 1)	(4, 3)	(4, 3)	(4, 3)	(4, 3)	(4, 3)
400	0.05	(0, 0)	(0, 0)	(0, 1)	(0, 1)	(0, 1)	(4, 3)	(4, 3)	(4, 3)	(4, 3)	(4, 3)
200	0.1	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(4, 3)	(0, 3)	(4, 3)	(4, 3)	(4, 3)
300	0.1	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 3)	(0, 3)	(4, 3)	(4, 3)	(4, 3)
400	0.1	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 3)	(0, 3)	(0, 3)	(4, 3)	(4, 3)

Table 6
Estimated numbers of row and column factors for different digit numbers.

Projection	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
No	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 4)
Yes	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)	(4, 5)

Table 7
Estimated numbers of row and column factors (or total number for vectorised methods) using different approaches in the literature, for different digit numbers. k_{max} is set as 10 for matrix-factor-model based approaches and 10^2 for vectorised methods.

	IterER	α-PCA	TensorTS				Vectorised		
			iTIP-IC	iTOP-IC	iTIP-ER	iTOP-ER	TCorTh	ER	IC
$\hat{k}_1$	4	4	6	6	4	4	8	4	12
$\hat{k}_2$	5	1	5	6	3	5	6		

almost always (save for some exceptions, based on a large α and a high threshold $f(S)$) estimated as $\widehat{k}_1 = 4$; indeed, the fact that results may be sensitive to $f(S)$ can be expected.[8] From an applied perspective, in the light of such sensitivity to $f(S)$, we would recommend to run the procedure with different specifications for $f(S)$, and then use the majority vote, as we did in our case.

Turning to k_2, all results indicate that this is not smaller than 3, and the most conservative approach (based on using $\alpha = 0.01$) indicates the possibility of having $\widehat{k}_2 = 5$. This may suggest that two common factors are less pervasive than the others. In order to avoid underestimation, we recommend taking $\widehat{k}_1 = 4$ and $\widehat{k}_2 = 5$ in this example. Indeed, in any real applications, we suggest the readers to try different combinations of α and threshold, and select the numbers of factors based on the real tolerance of underestimation and overestimation errors. Smaller α and threshold are in favour of H_0, but in higher risk of overestimation. Larger α and threshold will lead to opposite results.

For this example, we further compare the results for different digit numbers in Table 6 using only a small part of the images associated with a specific number. In this table, we report results corresponding to $S = 400$, $\alpha = 0.01$ and $f_5(S)$; we point out however that using different specifications leaves the results virtually unchanged. Results are remarkably stable across the different digits.

Finally, similarly to the previous application, we compare our results against those obtained using alternative criteria. The results in Table 7 show that results are broadly similar across the various techniques, in a more evident way than in the case of the previous exercise. In particular, all criteria indicate $\widehat{k}_1 \geq 4$. The iterative procedures by Han et al. (2022) show the same pattern as before, with Information Criteria having a tendency to estimate a larger number of common factors than the Eigenvalue Ratio statistic. Interestingly, the criteria proposed by Lam (2021) seem to overstate the number of common factors — this is particularly evident when comparing $\widehat{k}_1$, which the majority of criteria finds to be equal to 4, and it is found to be equal to 8 using the estimator by Lam (2021). As far as $\widehat{k}_2$ is concerned, the consensus is that $\widehat{k}_2 \geq 3$, with the majority vote agreeing with our estimate that $\widehat{k}_2 = 5$.

5.2. Multinational macroeconomic indices

Inspired by Chen and Fan (2021), we investigate the presence (and dimension) of a matrix factor structure in a time series of macroeconomic indices. In our application, we use the dataset employed by Yu et al. (2022), containing records of $p_2 = 10$ macroeconomic indices across $p_1 = 8$ OECD countries over $T = 130$ quarters, ranging from 1988Q1 to 2020Q2.

[8] By analogy, this case can be compared with reaching different conclusions when the significance level of a test is varied.

Table 8
Testing the null hypothesis $H_{01}: k_1 \geq 1$ for the macroeconomic indices dataset. Tests are based on $\alpha = 0.01, M = 100$, using the following thresholds: $f_1(S) = 1-\alpha-\sqrt{2\ln S/S}$, $f_2(S) = 1-\alpha-S^{-1/3}$, $f_3(S) = 1-\alpha-S^{-1/4}$, $f_4(S) = 1-\alpha-S^{-1/5}$, $f_5(S) = (1-\alpha)/2$.

S	No-projection					Projection				
	$f_1(S)$	$f_2(S)$	$f_3(S)$	$f_4(S)$	$f_5(S)$	$f_1(S)$	$f_2(S)$	$f_3(S)$	$f_4(S)$	$f_5(S)$
200	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept
300	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept
400	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept

Table 9
Testing the null hypothesis $H_{01}: k_2 \geq 1$ for the macroeconomic indices dataset. Tests are based on $\alpha = 0.01, M = 100$, using the following thresholds: $f_1(S) = 1-\alpha-\sqrt{2\ln S/S}$, $f_2(S) = 1-\alpha-S^{-1/3}$, $f_3(S) = 1-\alpha-S^{-1/4}$, $f_4(S) = 1-\alpha-S^{-1/5}$, $f_5(S) = (1-\alpha)/2$.

S	No-projection					Projection				
	$f_1(S)$	$f_2(S)$	$f_3(S)$	$f_4(S)$	$f_5(S)$	$f_1(S)$	$f_2(S)$	$f_3(S)$	$f_4(S)$	$f_5(S)$
200	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept
300	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept
400	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept

Table 10
Estimated numbers of row and column factors for the macroeconomic indices dataset. Five ways to select the threshold: $f_1(S) = 1-\alpha-\sqrt{2\ln S/S}$, $f_2(S) = 1-\alpha-S^{-1/3}$, $f_3(S) = 1-\alpha-S^{-1/4}$, $f_4(S) = 1-\alpha-S^{-1/5}$, $f_5(S) = (1-\alpha)/2$.

S	No-projection				
	$f_1(S)$	$f_2(S)$	$f_3(S)$	$f_4(S)$	$f_5(S)$
200	(1, 4)	(1, 4)	(1, 4)	(1, 4)	(2, 4)
300	(1, 3)	(1, 0)	(1, 4)	(1, 4)	(2, 4)
400	(1, 3)	(1, 3)	(1, 4)	(1, 4)	(2, 4)
S	Projection				
	$f_1(S)$	$f_2(S)$	$f_3(S)$	$f_4(S)$	$f_5(S)$
200	(1, 4)	(1, 3)	(1, 4)	(2, 4)	(2, 4)
300	(1, 3)	(1, 3)	(1, 3)	(1, 4)	(2, 4)
400	(1, 3)	(1, 3)	(1, 3)	(1, 4)	(2, 4)

Whilst we refer to Yu et al. (2022) for details, the countries are the United States, the United Kingdom, Canada, France, Germany, Norway, Australia and New Zealand, which can be naturally divided into three groups as North American, European and Oceania based on their geographical locations. The indices are from 4 major groups, namely consumer price, interest rate, production, and international trade.[9] As in Yu et al. (2022), we use the log-differences of each index, and each series is standardised.

We begin with testing whether there exists a matrix factor structure in the data. Results are in Tables 8 and 9. This result is similar to Chen and Fan (2021): if the factor structure has a Kronecker-product structure, then we cannot reject the null that there is a *two-way* factor structure, i.e. $k_1, k_2 \geq 1$.

We now turn to determining the dimensions of the row and column factor spaces k_1 and k_2. The empirical exercise in Chen and Fan (2021) demonstrates that, possibly owing to the small cross-sectional sizes, the estimated number of common factors differs considerably depending on the estimation method employed. Table 10 reports the estimated numbers of common factors in the form of (a, b) where a denotes the number of common row factors, and b the number of column factors. Using the results in Section 4 as guidelines (see in particular Section B.3.2), we have used $M = 150$, $S = 200$ and $\alpha = 0.01$; by way of robustness check, we have also considered different values of α, noting that, as α increases, our proposed procedure is more and more in favour of rejecting the existence of factors, and we have changed M to $M = 250$, obtaining the same results as reported here.[10] Here, we report results using different combinations of S and $f(S)$ to shed further light on the impact of these specifications; in particular, we use: the thresholds employed also in Section 4 – i.e. $f(S) = S^{-a}$ with $a = 1/3$, $1/4$ and $1/5$; a conservative threshold, $f(S) = \sqrt{2\ln S/S}$; and a very "liberal" one, with $f(S) = (1-\alpha)/2$.

Results are only partly affected by the choice of S and $f(S)$, which play a very minor role (a desirable form of robustness). As pointed out in Section 4, the projection technique should work better in finite samples, but in our

[9] In particular, we have considered the following indices, grouped by family: productivity (comprising: Total Index excluding Construction, Total Manufacturing, and GDP), CPI (comprising Food, Energy, and CPI Total), interest rates (long-term government bond yields, and 3-month Interbank rates and yields), and international trade (comprising total exports and total imports, both measured by value).

[10] Unreported results show that, for $\alpha \geq 0.05$, one would find $k_1 = 1$ and $k_2 = 0$, thus rejecting a matrix factor structure altogether. This reinforces the findings in the previous section, where it was noted that our procedure, in small samples, requires a smaller α in order to estimate the factor dimensions correctly.

Table 11
Estimated numbers of row and column factors (or total number for vectorised methods) using different approaches in the literature, for the macroeconomic indices dataset. k_{max} is set as 6 for matrix-factor-model based approaches and 6^2 for vectorised methods.

	IterER	α-PCA	TensorTS				TCorTh	Vectorised	
			iTIP-IC	iTOP-IC	iTIP-ER	iTOP-ER		ER	IC
$\hat{k}_1$	1	1	1	1	2	1	1	1	2
$\hat{k}_2$	5	2	1	1	2	2	3		

application results are actually comparable between the two techniques. According to Table 10, the number of row factors is at most $k_1 = 2$: whilst there is strong evidence in favour of at least one common factor (thus confirming that there is a matrix factor structure, as also found by Chen and Fan (2021) using the eigenvalue ratio approach), the second factor seems to be weaker, and deciding whether $\widehat{k}_1 = 1$ or 2 can be done on account of the researcher's preference for (possible) underestimation versus overestimation. Reading these results in conjunction with Table 9 in Yu et al. (2022) would suggest choosing $\widehat{k}_1 = 2$: factors broadly represent the different geographical locations, but European countries (particularly the largest economy, Germany) seem to also share a common factor structure with North America, speaking to the integration between the two economic areas. As far as $\widehat{k}_2$ is concerned, using the majority vote when applying the projection technique suggests $\widehat{k}_2 = 4$; even in this case, there seems to be some evidence in favour of $\widehat{k}_2 = 3$ also, again suggesting that, possibly, the fourth common factor is weaker than the others. Interestingly, the results in Chen and Fan (2021) using two different techniques (respectively, a scree-plot and an eigenvalue ratio approach) indicate that k_2 may range between 2 and 6, so our proposed approach offers a considerable refinement; Yu et al. (2022) also find $\widehat{k}_2 = 4$ or 5, but their results with $\widehat{k}_2 = 4$ show that this estimate explains the data very well, and it matches the four groups to which the indices belong which is an intuitive and meaningful finding.

Finally, by way of comparison we report the estimated values of the number of row and column common factors using various techniques available in the literature; results are in Table 11. There is broad consensus across all techniques as far as $\widehat{k}_1$ is concerned — one common factor is found by virtually all criteria, with the exception of the iTIP-ER criterion which indicates $\widehat{k}_1 = 2$. This is in line with our estimates, which (as mentioned above) suggest the presence of one strong common factor, and also the possible presence of a (weaker) second common factor. Conversely, there seems to be less consensus when estimating k_2. All criteria indicate a small value of $\widehat{k}_2$, which, also in the light of the empirical exercise in Yu et al. (2022), seems to be an understatement of the true number of common factors. As mentioned above, the α-PCA criterion delivers very different values of $\widehat{k}_2$ depending on the value of α (in our case, we have used $\alpha = 0$ as in the Monte Carlo exercise); the iterative procedures by Han et al. (2022) indicate that $\widehat{k}_2 = 1$ or 2, thus confirming the findings in Table B.6 which suggest a tendency to understate the number of common factors in small samples when this is larger than one. The criterion by Lam (2021), on the other hand, is the one closest to our findings, indicating that $\widehat{k}_2 = 3$. As a final remark, we point out that using the *IC* criterion suggests that the total number of factors $k = k_1 \times k_2$ is equal to 2. This result does not contradict the presence of a matrix structure in the data, but it indicates that fewer factors may be present along the rows and columns. On the other hand, we cannot rule out the possibility that some common factors may be weak, and not picked by *IC* (see the simulations in Uematsu and Yamagata, 2021 and Uematsu and Yamagata, 2022).

6. Conclusions and discussions

In this contribution, we studied the important issue of determining the presence and dimension of the row and column factor structures in a series of matrix-valued data exhibiting a Kronecker product structure in the loadings. Our methodology allows to check whether there is a factor structure in either dimension (row and column), thus helping the researcher decide whether data should be studied using the techniques developed by the literature for a standard vector factor model, or whether different techniques should be employed that are specific to tensor-valued data. In addition to finding evidence of a factor structure, we also proposed a methodology to estimate the numbers of common row and column factors.

Technically, our methodology is based on exploiting the eigen-gap which is found, in the presence of common factors, in the sample second moment matrix of the series. For each eigenvalue, we propose a test for the null that it diverges (as opposed to being bounded). Our tests are similar to the randomised tests (designed for vector factor models) proposed in Trapani (2018). However, we substantially refine rates via a different method of proof, and (crucially) we propose a "strong", Law-of-the-Iterated-Logarithm-inspired, decision rule which does away with the randomness, thus ensuring that all researchers using the same datasets will obtain the same results. In our paper, we proposed two procedures, based on two different ways of computing the sample second moment matrix: specifically, we use a "flattened" version of the matrix-valued series, and a projected version thereof, as proposed in Yu et al. (2022). We found that both techniques work very well in large samples, but, in small samples, the projection-based method is superior in all scenarios considered, also outperforming other existing methods.

Several important issues remain outstanding. In particular, from the outset, we have assumed that model (1.3) is correct, i.e. that the loading space has a Kronecker product structure. As we discussed in the introduction, under this

assumption the separate estimation of the loadings matrices C and R is advantageous since it entails a substantial dimensionality reduction: under (1.3), the estimation of $p_1k_1 + p_2k_2$ coefficients is required, compared to estimating Λ in (1.1), which contains $p_1p_2k_1k_2$ coefficients. Moreover, C and R have a clear interpretation, and estimating them allows to understand the interplay between the row factors and column factors, whereas Λ does not allow for such an interpretation. However, all these advantages are predicated on (1.3) being correctly specified in the first place. If this is the case, it would be possible to construct some pathological counterexamples in which (1.1) is correct, whereas (1.3) is not, and – when mistakenly using (1.3) and the techniques proposed in this paper – our tests find $k_1 = 0$ and $k_2 = 0$ even when the dimension of the factor space in (1.1), k, is strictly positive.[11] Whilst this issue goes beyond the scope of the present paper, we offer a more in-depth discussion of this issue through an example. Consider the case of $p_2 < p_1$ and consider the following vector factor model

$$\text{Vec}\,(X_t) = \Lambda F_t + \text{Vec}\,(E_t)\,. \tag{6.1}$$

We assume that the loadings Λ satisfy

$$\Lambda = (\Lambda_1', \ldots, \Lambda_{p_2}')', \text{ satisfying } \Lambda_i'\Lambda_j/p_1 = 0 \text{ when } i \neq j, \text{ and } \Lambda_i'\Lambda_i/p_1 = I_k;$$

that is, the blocks $\Lambda_1, \ldots, \Lambda_{p_2}$ of Λ are orthonormal, and their columns span a p_2-dimensional linear space. In this setting, it holds that $\Lambda'\Lambda/(p_1p_2) = I_k$, which satisfies the strong/pervasive factor condition for the vector factor model. Eq. (6.1) can be rewritten artificially (i.e., without meaningful row and column cross-sections) in matrix form, viz.

$$X_t = (\Lambda_1 F_t, \ldots, \Lambda_{p_2} F_t) + E_t, \tag{6.2}$$

where the errors are such that $E\left(E_t'E_t\right)/p_1$ has bounded eigenvalues, as also stipulated by our Assumption B3; we assume for simplicity that $E\left(F_t'F_t\right) = c_0$. Then, by standard algebra, it follows that the column covariance matrix is given by

$$\Sigma_c = E(\Lambda_1 F_t, \ldots, \Lambda_{p_2} F_t)'(\Lambda_1 F_t, \ldots, \Lambda_{p_2} F_t)/p_1 = c_0 I_{p_2}.$$

Thus, the second moment matrix of the signal $(\Lambda_1 F_t, \ldots, \Lambda_{p_2} F_t)$ has eigenvalues of the same order of magnitude as the second moment matrix of the idiosyncratic errors, which entails that there are no strong (or even weak) column factors. Similarly, considering the row covariance matrix and assuming, again for simplicity, that $E\left(F_tF_t'\right) = c_1 I_k$, it holds that

$$\Sigma_r = E(\Lambda_1 F_t, \ldots, \Lambda_{p_2} F_t)(\Lambda_1 F_t, \ldots, \Lambda_{p_2} F_t)'/p_1 = c_1 \left(\sum_{i=1}^{p_2} \Lambda_i\Lambda_i'\right)/p_1.$$

It is easy to see that this matrix is idempotent,[12] and therefore its eigenvalues belong in $\{0, 1\}$; hence, the row covariance matrix $p_1\Sigma_r/p_2$ has eigenvalues 0 and p_1/p_2. If p_1/p_2 is bounded, the largest eigenvalues of the row signal matrix Σ_r are all bounded, which again leads to finding no strong (or even weak) row factors. Thus, we conclude that a methodology based on assuming (1.3) incorrectly detects no common factors, either in the columns or in the rows, despite the existence of a factor structure in the vector factor model. In Section B.3.6 in the Supplement, we also report some Monte Carlo evidence when the Kronecker product structure is not satisfied, showing that our methodology is robust under "local" mis-specifications of the factor structure.

In general, as discussed in the introduction, (1.3) is likely to be an adequate model where there is a meaningful matrix structure, with economically meaningful row and column cross-sections; the examples discussed in the introduction are some of the possible cases in which (1.3) is a natural formulation, based on the very nature of the data. This said, as a preliminary step in the analysis, it would nonetheless be highly desirable to have a formal test to check whether a Kronecker product structure does exist in the loadings space Λ. We are not aware of any such test in a high-dimensional context. A recent contribution by Guggenberger et al. (2022) provides a test for the null of having a Kronecker product structure in a fixed dimensional matrix; similarly, Chen et al. (2021a) propose a test for the Kronecker product structure in the context of a Vector AutoRegressive model for a matrix-valued time series, but in that case the row and column dimensions of the data X_t are both fixed. The highly nontrivial problem of developing a test for a Kronecker product structure in the large dimensional case is currently under investigation by the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

11 We are grateful to an anonymous Referee for pointing this out to us.

12 Indeed, it holds that

$$\left(\frac{1}{p_1}\sum_{i=1}^{p_2}\Lambda_i\Lambda_i'\right)\left(\frac{1}{p_1}\sum_{i=1}^{p_2}\Lambda_i\Lambda_i'\right) = \frac{1}{p_1^2}\sum_{i,j=1}^{p_2}\Lambda_i\Lambda_i'\Lambda_j\Lambda_j' = \frac{1}{p_1^2}\sum_{i=1}^{p_2}\Lambda_i\Lambda_i'\Lambda_i\Lambda_i' = \frac{1}{p_1}\sum_{i=1}^{p_2}\Lambda_i\Lambda_i';$$

Acknowledgements

He's work is supported by NSF China (12171282,11801316), National Statistical Scientific Research Key Project (2021LZ09), Project funded by China Postdoctoral Science Foundation (2021M701997) and the Fundamental Research Funds of Shandong University, Young Scholars Program of Shandong University, China. Kong's work is partially supported by NSF China (71971118 and 11831008) and the WRJH-QNBJ Project and Qinglan Project of Jiangsu Province. The authors would like to thank the Editor Xiaohong Chen, an anonymous Associate Editor, and three anonymous Referees, whose helpful comments have greatly improved the quality and focus of the paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2023.02.008.

Further discussions on the assumptions, the technical proofs of the main results and extra simulation and empirical studies are included in the Supplementary Material.

References

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.
Ando, T., Bai, J., 2016. Panel data models with grouped factor structure under unknown group membership. J. Appl. Econometrics 31 (1), 163–191.
Andreou, E., Gagliardini, P., Ghysels, E., Rubin, M., 2019. Inference in group factor models with an application to mixed-frequency data. Econometrica 87 (4), 1267–1305.
Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.
Bai, J., Wang, P., 2016. Econometric analysis of large factor models. Annu. Rev. Econ. 8, 53–80.
Cai, C., Chen, R., Xiao, H., 2019. Kopa: Automated Kronecker product approximation. arXiv:1912.02392.
Chen, E.Y., Chen, R., 2020. Modeling dynamic transport network with matrix factor models: With an application to international trade flow. arXiv:1901.00769.
Chen, E.Y., Fan, J., 2021. Statistical inference for high-dimensional matrix-variate factor models. J. Amer. Statist. Assoc. 1–18.
Chen, W., Lam, C., 2022. Rank and factor loadings estimation in time series tensor factor model by pre-averaging. arXiv:2208.04012.
Chen, E.Y., Tsay, R.S., Chen, R., 2020a. Constrained factor models for high-dimensional matrix-variate time series. J. Amer. Statist. Assoc. 115 (530), 775–793.
Chen, E.Y., Xia, D., Cai, C., Fan, J., 2020b. Semiparametric tensor factor analysis by iteratively projected SVD. arXiv:2007.02404.
Chen, R., Xiao, H., Yang, D., 2021a. Autoregressive models for matrix-valued time series. J. Econometrics 222 (1), 539–560.
Chen, X., Yang, D., Xu, Y., Xia, Y., Wang, D., Shen, H., 2021b. Testing and support recovery of correlation structures for matrix-valued observations with an application to stock market data. J. Econometrics.
Chen, R., Yang, D., Zhang, C.H., 2022. Factor models for high-dimensional tensor time series. J. Amer. Statist. Assoc. 117 (537), 94–116.
Corradi, V., Swanson, N.R., 2006. The effects of data transformation on common cycle, cointegration, and unit root tests: Monte Carlo and a simple test. J. Econometrics 132, 195–229.
El Karoui, N., 2005. Recent results about the largest eigenvalue of random covariance matrices and statistical application. Acta Phys. Polonica Ser. B 36 (9), 2681.
Gao, Z., Tsay, R.S., 2021. A two-way transformed factor model for matrix-variate time series. Econom. Stat..
Gao, Z., Yuan, C., Jing, B., Wei, H., Guo, J., 2021. A two-way factor model for high-dimensional matrix data. arXiv:2103.07920.
Geman, S., 1980. A limit theorem for the norm of random matrices. Ann. Probab. 8 (2), 252–261.
Geyer, C.J., Meeden, G.D., 2005a. Fuzzy and randomized confidence intervals and p-values. Statist. Sci. 358–366.
Geyer, C.J., Meeden, G.D., 2005b. Rejoinder: Fuzzy and randomized confidence intervals and P-values. Statist. Sci. 20 (4), 384–387.
Guggenberger, P., Kleibergen, F., Mavroeidis, S., 2022. A test for Kronecker product structure covariance matrix. J. Econometrics.
Han, Y., Chen, R., Zhang, C.H., 2022. Rank determination in tensor factor model. Electron. J. Stat. 16 (1), 1726–1803.
Kong, X.B., Liu, Y.X., Yu, L., Zhao, P., 2022. Matrix quantile factor model. arXiv preprint arXiv:2208.08693.
Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer 42 (8), 30–37.
Lam, C., 2021. Rank Determination for Time Series Tensor Factor Model Using Correlation Thresholding. Technical Report, Working paper LSE.
Lam, C., Yao, Q., 2012. Factor modeling for high-dimensional time series: Inference for the number of factors. Ann. Statist. 40, 694–726.
Onatski, A., 2009. Testing hypotheses about the number of factors in large factor models. Econometrica 77, 1447–1479.
Rio, E., 1995. A maximal inequality and dependent Marcinkiewicz-Zygmund strong laws. Ann. Probab. 23 (2), 918–937.
Salzer, H.E., Zucker, R., Capuano, R., 1952. Table of the zeros and weight factors of the first twenty Hermite polynomials. J. Res. Natl. Bur. Stand. 48 (2), 111.
Shao, Q.M., 1995. Maximal inequalities for partial sums of ρ-mixing sequences. Ann. Probab. (23), 948–965.
Trapani, L., 2018. A randomized sequential procedure to determine the number of factors. J. Amer. Statist. Assoc. 113 (523), 1341–1349.
Uematsu, Y., Yamagata, T., 2021. Inference in sparsity-induced weak factor models. J. Bus. Econom. Statist. 1–14.
Uematsu, Y., Yamagata, T., 2022. Estimation of sparsity-induced weak factor models. J. Bus. Econom. Statist. 1–15.
Wang, D., Liu, X., Chen, R., 2019. Factor models for matrix-valued high-dimensional time series. J. Econometrics 208 (1), 231–248.
Wu, W.B., 2005. Nonlinear system theory: Another look at dependence. Proc. Natl. Acad. Sci. USA 102 (40), 14150–14154.
Yu, L., He, Y., Kong, X., Zhang, X., 2022. Projected estimation for large-dimensional matrix factor models. J. Econometrics 229 (1), 201–217.