WORKING PAPER

ESTIMATION OF NON-GAUSSIAN FACTORS USING HIGHER-ORDER MULTI-CUMULANTS IN WEAK FACTOR MODELS

Wanbo Lu
Guanglin Huang
Kris Boudt

March 2024
2024/1085

GHENT UNIVERSITY

Department of Economics

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

Estimation of non-Gaussian factors using higher-order multi-cumulants in weak factor models

Wanbo Lu[a], Guanglin Huang[b,d,*], Kris Boudt[c,d,e]

[a]*School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu 611130, China*
[b]*Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China*
[c]*Department of Economics, Universiteit Gent, 9000 Gent, Belgium*
[d]*Department of Business, Vrije Universiteit Brussel, 1050 Brussels, Belgium*
[e]*School of Business and Economics, Vrije Universiteit Amsterdam, 1081 Amsterdam, The Netherlands*

Abstract

We estimate the latent factors in high-dimensional non-Gaussian panel data using the eigenvalue decomposition of the product between the higher-order multi-cumulant and its transpose. The proposed Higher order multi-cumulant Factor Analysis (HFA) approach comprises an eigenvalue ratio test to select the number of non-Gaussian factors and uses the eigenvector to estimate the factor loadings. Unlike covariance-based approaches, HFA remains reliable for estimating the non-Gaussian factors in weak factor models with Gaussian error terms. Simulation results confirm that HFA estimators improve the accuracy of factor selection and estimation compared to covariance-based approaches. We illustrate the use of HFA to detect and estimate the factors for the FRED-MD data set and use them to forecast the monthly S&P 500 equity premium.

Keywords: Higher-order multi-cumulants, High-dimensional factor models, Weak factors, Consistency, Eigenvalues
JEL: G11, G12, G15

1. Introduction

Factor models are widely used to characterize the low dimensional common structure from large panels of economic data (Stock & Watson, 2002). Factor model approaches differ in how they estimate the latent structure. The most common approach is to use principal component analysis (PCA) (Bai, 2003; Bai & Ng, 2013). Recently, the Gaussian quasi-maximum likelihood estimator has been developed to reach a higher efficiency than the PCA estimator (Bai et al., 2012; Bai &

*Corresponding author: SWUFE, Liutai Avenue 555, 611130 Chengdu, China. Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. (Email: huanggl@swufe.edu.cn)

Li, 2016). However, as mentioned in De Mol et al. (2008) and Onatski (2012), when the factors have weak influential power, the PCA estimators have a slow convergence rate and can even be inconsistent. This poor performance is due to the low signal-to-noise ratio of the factors when their explanatory power decreases. As mentioned in Fan et al. (2022), the selection of common factors based on the difference between the eigenvalues of the covariance (correlation) matrix becomes inconsistent when the factors' explanatory power is smaller than a threshold value, which they refer to as the minimal signal strength.

In this study, we propose to exploit the information in the higher-order multi-cumulants to characterize the dependence structure when the data have underlying weak non-Gaussian factors. Under the proposed framework, we assume that the non-Gaussianity of the observed variables only comes from the factors and set up a linear factor model with non-Gaussian factors and Gaussian idiosyncratic errors. We use an eigenvalue-based approach to conduct factor analysis with $(N, T) \to \infty$, labeled as Higher-order multi-cumulant Factor Analysis (HFA). We contribute to the extant literature in two ways. First, we propose an eigenvalue ratio-based test for determining the number of non-Gaussian factors consistently, especially the weak factor case. Second, we present a computationally convenient factor estimation approach based on eigenvalue decomposition. We study the asymptotic properties of the HFA estimators and prove the efficiency gain of HFA estimators compared to PCA estimators on non-Gaussian factors.

To determine the number of non-Gaussian factors, we develop a statistical test based on the eigenvalues of the product between the higher-order multi-cumulant and its transpose. A few recent studies have considered the problem of determining the number of factors for the modeling of higher-order multi-cumulants. However, this research differs from these studies significantly. Jondeau et al. (2018) use a threshold method to identify the factors that drive co-skewness and co-kurtosis structures across a large set of time series. Boudt et al. (2020) suggests using the information criteria AIC or BIC based on the nearest-distance estimation between the factor models implied by the higher-order co-moments and their sample version to select an appropriate factor model. Their framework assumes a strict factor model only suitable for a small N. Lu & Huang (2022) propose a higher-order cumulant test to identify the number of non-Gaussian factors in the observed factor model. We determine the number of non-Gaussian factors by selecting the number of singular values that drive the higher-order multi-cumulants. Ahn & Horenstein (2013) exploit the well-known fact that the R largest eigenvalues of the covariance matrix of N observed variables grow unboundedly as N increases to infinity, whereas the other eigenvalues remain bounded. A similar property holds for the higher-order multi-cumulants — the R largest singular values of the k-th order multi-cumulant tensor are unbounded when N increases to infinity, but the others remain bounded.

Interestingly, this property in higher-order multi-cumulants holds in the weak factor model but disappears in the covariance matrix. We exploit this to estimate the number of factors and show how this differs from estimating R by using the covariance matrix. For asymmetric distributions, the third-order test is sufficient to estimate R. For symmetric non-Gaussian distributions, we need the fourth-order test to obtain a consistent estimate of R. The proposed Generalized Eigenvalue Ratio (GER) estimator is obtained by maximizing the ratio of two adjacent singular values of the higher-order sample multi-cumulant arranged in descending order. We show the consistency of the proposed factor number estimator with $(N, T) \to \infty$. Our extensive simulation results indicate that the GER estimator performs adequately in finite samples.

Further, we contribute to the literature on non-Gaussian factor estimation. We propose a tractable eigenvalue-based approach to estimate factors and loadings. The HFA approach guarantees the consistency and asymptotic normality of the estimated non-Gaussian factors and factor loadings in the weak factor model. It is easy to implement because it only depends on the eigenvalue decomposition of the product between the higher-order multi-cumulant and its transpose. The proposed HFA framework estimates latent factors based on the k-th order cumulants. When $k = 2$, HFA coincides with PCA. In the case of weak non-Gaussian factors, we recommend considering the higher order cumulants ($k = 3, 4$). Thus, the proposed HFA approach nests PCA as a special case. An alternative approach to extract non-Gaussian factors is the independent component analysis (ICA), which is based on minimizing the statistical dependence between factors (see e.g. Cardoso & Souloumiac, 1993; Bonhomme & Robin, 2009). It is widely used in signal processing, image processing, and neural recognition (Bonhomme & Robin, 2009). In high-dimensional ICA, PCA is usually applied prior to classic ICA (Risk et al., 2019). Consequently, the ICA estimator is a rotation of the PCA estimator. Therefore, ICA does not improve the efficiency for estimating the factors and factor loadings.

The remainder of this paper is organized as follows. Section 2 provides the factor model assumed in HFA and the notations, and Section 3 presents the tests to determine the number of factors. Further, Section 4 introduces the HFA estimates and their properties, and Section 5 gives the prediction framework using an HFA factor-augmented regression. Subsequently, Section 6 reports our simulation experiments and findings, and Section 7 presents an application to market premium prediction. Finally, Section 8 provides the concluding remarks. Furthermore, a Supplementary Appendix provides additional details about this study, including several alternative factor estimation and selection approaches, computational issues, HFA estimates in the presence of Gaussian factors, more robustness checks, and the R code for the HFA estimator, which is available publicly in the `hofa` package.

2. Factor model

Throughout the paper, we use $\sigma_r(A)$ to denote the r-th largest singular value of real value matrix A and $\psi_r(B)$ to represent the r-th largest eigenvalue of positive semi-definite matrix B. The Frobenius norm of a matrix A is denoted by $\|A\| = \sqrt{\text{tr}(A'A)}$, where $\text{tr}(\cdot)$ is the trace of matrix. We use the notation $a \asymp b$ to denote $a = O(b)$ and $b = O(a)$. We use $a \ll b$ to denote $a = o(b)$ and $a \gg b$ to denote $b = o(a)$. $\lesssim$ (or $\gtrsim$) to present $\leq$ (or $\geq$) up to a positive constant. $A^{\otimes k} = A\underbrace{\otimes \ldots \otimes}_{k} A$, where $\otimes$ is the Kronecker product. $[z]$ is the integer part of a real number z.

2.1. The model

This subsection introduces the main model that will be used throughout the paper. Let x_{it} be the observed variable for $i = 1, \ldots, N$ cross-sectional units at time $t = 1, \ldots, T$. Assume that the observed data are generated by an $R \times 1$ vector of common factors, $f_t = [f_{1t}, f_{2t}, \ldots, f_{Rt}]'$. The factor model is as follows:

$$x_{it} = \lambda_i' f_t + e_{it}, \tag{1}$$

where $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{iR})'$ is the $R \times 1$ vector of factor loading for variable i, and e_{it} represents the idiosyncratic components of variable i at time t. The factors, factor loadings, and idiosyncratic errors are not observable. For convenience, the time series model (1) can be written as complete panel data:

$$X = F\Lambda' + E, \tag{2}$$

where $X = (x_1, \ldots, x_i, \ldots, x_N)$ and $x_i = (x_{i1}, \ldots, x_{iT})$, $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)'$, $F = (f_1, f_2, \ldots, f_T)'$, and $E = (e_1, \ldots, e_i, \ldots, e_N)$ and $e_i = (e_{i1}, \ldots, e_{iT})$. Following Bai & Ng (2002), we treat the entries in Λ as parameters and those in F as random variables.

Unlike the existing literature, we use the information of higher-order multi-cumulants — instead of the covariance matrix — to identify the factor structure. Further, we study the properties of the eigenvalue decomposition of the product of the higher-order multi-cumulant and its transpose. Subsequently, we find the efficiency of this method in estimating weak factor models. We first give a short introduction of the higher-order multi-cumulant and its singular values and eigenvalues. Working with multi-cumulants is convenient in a linear factor model with independence assumptions between f_t and e_t. For instance, the multi-cumulants for the second-, third-, and fourth-order

moments of $z_t = (z_{1t}, z_{2t}, \ldots, z_{Qt}) \in \mathbb{R}^{Q\times 1}$ with zero means are

$$\begin{aligned}
\kappa_{z,i_1i_2,t} &= m_{z,i_1i_2,t}, \\
\kappa_{z,i_1i_2i_3,t} &= m_{z,i_1i_2i_3,t}, \\
\kappa_{z,i_1i_2i_3i_4,t} &= m_{z,i_1i_2i_3i_4,t} - m_{z,i_1i_2,t}m_{z,i_3i_4,t} - m_{z,i_1i_3,t}m_{z,i_2i_4,t} - m_{z,i_1i_4,t}m_{z,i_2i_3,t},
\end{aligned} \tag{3}$$

where $\kappa_{z,i_1i_2\ldots i_k,t}$ is the k-th order multi-cumulant of $z_{i_1t}, z_{i_2t}, \ldots, z_{i_kt}$, $i_q \in \{1, 2, \ldots, Q\}$ for $1 \leq q \leq k$ and $k = 2, 3, 4$, and $m_{z,i_1i_2\ldots i_k,t} = \mathbb{E}(z_{i_1t}z_{i_2t}\ldots z_{i_kt})$. Following Kolda & Bader (2009), we can express the k-th order multi-cumulant tensor of z_t in a matrix form:

$$\mathbf{C}_{z,t}^{(k)} = \{\kappa_{z,i_1j,t}\} \in \mathbb{R}^{Q\times Q^{k-1}}, \tag{4}$$

where $j = 1 + \sum_{p=2}^{k}(i_p - 1)Q^{p-2}$. We define

$$\mathbf{C}_{z}^{(k)} \equiv \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbf{C}_{z,t}^{(k)}, \tag{5}$$

if the constant matrix $\mathbf{C}_z^{(k)}$ exists. The matrix $\mathbf{C}_z^{(k)}$ is labeled as the (population) k-th order multi-cumulant for the sequence $\{z_t\}_{t=1}^T$ as $T \to \infty$.

Furthermore, with the available observations $\{z_t\}_{t=1}^T$, we define

$$\widetilde{\mathbf{C}}_{z}^{(k)} \equiv \frac{1}{T}\sum_{t=1}^{T} \widetilde{\mathbf{C}}_{z,t}^{(k)}. \tag{6}$$

where $\widetilde{\mathbf{C}}_{z,t}^{(k)}$ is defined similarly by using the notations in (3) and (4) and replacing $m_{z,i_1i_2\ldots i_k,t}$ with $\widetilde{m}_{z,i_1i_2\ldots i_k,t} = z_{i_1t}z_{i_2t}\ldots z_{i_kt}$. The matrix $\widetilde{\mathbf{C}}_z^{(k)}$ is labeled as the sample k-th order multi-cumulant for the sequence $\{z_t\}_{t=1}^T$. In this paper, z_t represents either x_t, f_t, or e_t.

In this study, we propose to identify the factors and loadings by performing HFA on the higher-order multi-cumulant matrix instead of PCA on the sample covariance matrix. Now, for the factors in HFA to be identified successfully, we need to make some assumptions on the higher-order moments of the factors as follows:

ASSUMPTION A: The factors

(i) $\mathbb{E}\|f_t\|^{2K} < C_1$ for a real integer $K \geq 3$, where C_1 is a finite constant.

(ii) For the sequence $\{f_t\}_{t=1}^T$, its k-th order sample multi-cumulant ($2 \leq k \leq K$) defined in (6) satisfies $\widetilde{\mathbf{C}}_f^{(k)} \xrightarrow{p} \mathbf{C}_f^{(k)}$ as $T \to \infty$, where $\mathbf{C}_f^{(k)}$ is a constant matrix defined in (5). Let

$\phi_j^{(k)} = \sigma_j(\mathbf{C}_f^{(k)})$, $0 < \phi_j^{(k)} < \infty$ for $2 \leq k \leq K$ and $j = 1, 2, \ldots, R$.
(iii) The variables f_t and e_t are mutually independent.

Assumptions A(i) is common in factor analysis. Assumptions A(ii) require the sample k-th order multi-cumulant of the sequence $\{f_t\}_{t=1}^T$ converge to a constant matrix as $T \to \infty$, which is also common in static factor model when $k = 2$, see e.g. Bai (2003) and Ahn & Horenstein (2013). Moreover, we require all the singular value of $\mathbf{C}_f^{(k)}$ are nonzero for $k \geq 3$, which adds a non-Gaussian feature to the factors. Therefore, we refer to the factors satisfying Assumption A(ii) as "non-Gaussian factors". The non-Gaussian factor assumption is common in Independent Component Analysis (see in Cardoso & Souloumiac, 1993; Bonhomme & Robin, 2009). When Gaussian factors exist, we propose a two-stage procedure in the Supplementary Appendix to solve this problem. If the Gaussian factors and non-Gaussian factors are mutually dependent, they can be identified directly because the matrix $\mathbf{C}_f^{(k)}$ is still of full rank. Specifically , let $f_{1t} \sim i.i.d.\mathcal{N}(0,1)$ and $f_{2t} = (f_{1t})^2$. The third-order multi-cumulant $\mathbf{C}_f^{(3)}$ has the following form:

$$\mathbf{C}_f^{(3)} = \begin{pmatrix} \mathbb{E}[f_{1t}^3] & \mathbb{E}[f_{1t}^2 f_{2t}] & \mathbb{E}[f_{1t}^2 f_{2t}] & \mathbb{E}[f_{1t} f_{2t}^2] \\ \mathbb{E}[f_{1t}^2 f_{2t}] & \mathbb{E}[f_{1t} f_{2t}^2] & \mathbb{E}[f_{1t} f_{2t}^2] & \mathbb{E}[f_{2t}^3] \end{pmatrix} = \begin{pmatrix} 0 & 3 & 3 & 0 \\ 3 & 0 & 0 & 15 \end{pmatrix}$$

The singular values $\sigma_1(\mathbf{C}_f^{(3)}) = \sqrt{234}$ and $\sigma_2(\mathbf{C}_f^{(3)}) = \sqrt{18}$ are nonzero; therefore, Assumption A(ii) still holds. Nonlinear dependence between the Gaussian and non-Gaussian factors also provides singular values of $\mathbf{C}_f^{(k)}$, only when the factors are Gaussian and independent of non-Gaussian factors, then $\mathbf{C}_f^{(k)}$ is not full rank. Therefore, we cannot use the higher-order multi-cumulant to identify them. Overall, Assumption A(ii) is required for the HFA framework. However, when Assumption A(ii) is violated, we need a two-stage procedure to estimate all factors in the presence of independent Gaussian factors, see in the Supplementary Appendix. Assumption A(iii) requires exogenous idiosyncratic errors. This assumption is common in latent factor modelling (see, for example, Bonhomme & Robin, 2009; Boudt et al., 2020; Chen et al., 2021). The assumption is needed to derive the theoretical properties of the HFA approach.

For factor loadings, we treat them as parameters like Bai (2003) and Ahn & Horenstein (2013). The assumptions on factor loadings are as follows:

ASSUMPTION B: The factor loadings
(i) $\|\lambda_i\| < C_2$ for all $i = 1, 2, \ldots, N$ with a finite constant C_2.
(ii) $\lim_{N\to\infty} \frac{1}{N^{1-\alpha}} \Lambda'\Lambda = \Sigma_\Lambda$ for some constant $\alpha \in [0, 1]$.

Assumption B requires that Λ is column full rank and that $\Lambda'\Lambda/N^{1-\alpha}$ converges to a positive definite matrix. Following the assumptions of De Mol et al. (2008), Onatski (2012), Bailey et al.

(2021), and Freyaldenhoven (2022), we allow the eigenvalues of $\Lambda'\Lambda$ to diverge at rate $N^{1-\alpha}$, $\alpha \in [0,1]$.[1] Standard strong factor models typically assume that the eigenvalues of $\Lambda'\Lambda$ diverge at rate N and the largest eigenvalue of the covariance matrix of error terms e_{it} are bounded by a finite constant. If $\alpha > 0$, we describe the factor model as a weak factor case. Onatski (2012) studies the special case of $\alpha = 1$. Under this assumption, the eigenvalues of the covariance matrix of c_{it} ($c_{it} = \lambda_i' f_t$) are $O(1)$. We refer to this case as Onatski (2012)'s weak factor model.

The assumptions on the idiosyncratic errors are as follows:

ASSUMPTION C: The idiosyncratic errors

Let $E = (e_{it})_{T\times N} = UG_N^{1/2}$, where $U = (u_{it})_{T\times N}$. G_N is a $N \times N$ matrix, and $G_N^{1/2}$ is its symmetric square roots.

(i) For each i, $\{u_{it}\}_{t=1}^T$ is a strong mixing Gaussian sequence such that $\mathbb{E}(u_{it}) = 0$ and $\mathbb{E}(u_{it}^2) < \infty$. The mixing coefficients $\bar{\alpha}_i(\cdot)$ for $\{u_{it}\}_{t=1}^T$ satisfies $\max_i \bar{\alpha}_i(n) \leq C_{\bar{\alpha}}\tau^n$ for some $C_{\bar{\alpha}} > 0$ and $\tau \in (0,1)$. $\{u_{it}\}_{t=1}^T$ are independent across i.

(ii) $\sigma_1(G_N) < C_3$ uniformly in N with a finite constant C_3.

Assumption C models the idiosyncratic errors E as a weighted combination of the strong mixing sequence $\{u_{it}\}_{t=1}^T$. The specification allows accommodation for serial and cross-sectional correlation in the idiosyncratic errors. The strong mixing setting is common to characterize serial dependence of time series, see e.g. Su et al. (2016), Chang et al. (2018). The matrix G_N is used to characterize cross-sectional dependence.

Remark 2.1. *Assumption C(i) follows Su et al. (2016)'s Assumption A1(i). This strong mixing assumption is needed for all asymptotic results presented in this paper. Onatski (2010) and Ahn & Horenstein (2013) specify $E = R_T^{1/2}U^*G_N^{1/2}$, where $U^* = (u_{it}^*)_{T\times N}$ are i.i.d. standard normal random variables, $R_T^{1/2}$ and $G_N^{1/2}$ are the symmetric square root of $T \times T$ and $N \times N$ positive semidefinite matrices R_T and G_N, respectively. We thus have a different approach in describing the serial dependence which has the following properties: (i) we allow heterogeneity in the serial dependence of $\{e_{it}\}_{t=1}^T$ and characterize it using the mixing coefficient $\bar{\alpha}_i(\cdot)$; (ii) the serial dependence is characterized by $\max_i \bar{\alpha}_i(n) \leq C_{\bar{\alpha}}\tau^n$, which is more strict than the assumption $\sigma_1(R_T) < \infty$ in Onatski (2010) and Ahn & Horenstein (2013).*

The normality assumption on the error terms in Assumption C(i) implies different divergence rates of the singular values of the higher-order multi-cumulant of the factors and errors, which

[1] Generally, a factor can be weak in two ways: (i) it can affect all outcomes weakly as in De Mol et al. (2008) and Onatski (2012); (ii) it can affect only a subset of the outcomes as in Bailey et al. (2021) and Freyaldenhoven (2022).

is part of the conditions needed to show the consistency of the proposed HFA estimates even when the factor strength is weak. Throughout the paper, we use this assumption to derive the asymptotic results. However, it can be relaxed in the case of weak non-normal errors, which depend on the structure of G_N. We discuss it in subsection 4.3. Intuitively, the propagation of the potential non-normality of u_{it} on e_{it} is dampened because of the aggregation involved in its definition as $E = UG_N^{1/2}$. Bai (2003), Fan et al. (2022), and Freyaldenhoven (2022) account for the cross-sectional correlation on error terms, which implies $G_N^{1/2}$ is not a diagonal matrix. In the Supplementary Appendix, we show that when $\mathbb{E}[|u_{it}|^{2K}] < \infty$ for all i and the number of nonzero non-diagonal elements in $G_N^{1/2}$ diverges to infinity as $N \to \infty$, we have $e_t = G_N^{1/2} u_t$ being asymptotically normal according to the Lyapunov Central Limit Theorem (CLT) under mild assumptions on $G_N^{1/2}$. This is not satisfied in cases where $G_N^{1/2}$ has a limited number of non-zero elements, such as when $G_N^{1/2}$ is a block diagonal matrix with finite block size. By the Cauchy-Schwarz inequality, we can, also in such cases, still expect e_{it} to be nearer normal than u_{it} if u_{it} are independent and identically distributed, see for example in Theorem 3.1 of Granger (1976). If u_{it} are independent but not identically distributed, as mentioned in Theorem 5.1 of Granger (1976), we can expect e_{it} to be nearer normal than the least normal u_{it}. In the simulation study, we provide Monte Carlo evidence that the HFA approach remain reliable in the case of weak non-Gaussian errors if $G_N^{1/2}$ has a sufficiently large number of nonzero non-diagonal elements.

2.2. Identification conditions

In this subsection, we first derive the population covariance and higher-order multi-cumulants of x_t as implied by Assumptions A – C. Then we show that the columns of the factor loading matrix Λ in (2) are the eigenvectors to $\mathbf{C}_x^{(k)}\mathbf{C}_x^{(k)'}$ $(3 \leq k \leq K)$ up to a rotation matrix.

First, based on the factor model (1) and Assumption A and B, the k-th order multi-cumulant of $x_{i_1 t}, x_{i_2 t}, \ldots, x_{i_k t}$ $(3 \leq k \leq K)$ with $i_m \in \{1, 2, \ldots, N\}$ and $m \in \{1, \ldots, k\}$, can be expressed as

$$\kappa_{x, i_1 \ldots i_k, t} = \lambda_{i_1}' \mathbf{C}_{f,t}^{(k)} (\lambda_{i_2} \otimes \ldots \otimes \lambda_{i_k}) + \kappa_{e, i_1 \ldots i_k, t}, \tag{7}$$

where $\mathbf{C}_{f,t}^{(k)}$ is the k-th order multi-cumulant matrix of f_t defined in (4), and $\kappa_{e, i_1 \ldots i_k, t}$ is the k-th order multi-cumulant of $e_{i_1 t}, e_{i_2 t}, \ldots, e_{i_k t}$. By Assumption C, the idiosyncratic errors e_t are Gaussian distributed. It follows that $\kappa_{e, i_1 \ldots i_k, t} = 0$ for $i_m \in \{1, 2, \ldots, N\}$ and $m \in \{1, \ldots, k\}$. Therefore, we can rewrite (7) in matrix form as follows

$$\mathbf{C}_{x,t}^{(k)} = \Lambda \mathbf{C}_{f,t}^{(k)} (\Lambda'^{\otimes (k-1)}). \tag{8}$$

For a given N and $T \to \infty$, under Assumption A - C, the population k-th order multi-cumulant $\mathbf{C}_x^{(k)}$ $(3 \leq k \leq K)$ can be expressed as

$$\mathbf{C}_x^{(k)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{C}_{x,t}^{(k)} = \Lambda \mathbf{C}_f^{(k)} (\Lambda'^{\otimes(k-1)}), \tag{9}$$

where $\mathbf{C}_f^{(k)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{C}_{f,t}^{(k)}$ defined in (5) is the population k-th order multi-cumulant matrix of f_t.

It is well known that the factor F and the factor loading matrix Λ are not uniquely determined by (2), as we may replace (F, Λ) by $(F^*, \Lambda^*) = (FH'^{-1}, \Lambda H)$ for any invertible matrix H. Hence, as mentioned in Stock & Watson (2002), Bai (2003) and Bai & Ng (2013), we can only identify and estimate (F, Λ) up to a rotation matrix. To this end, we impose the normalization restrictions

$$\frac{1}{N}\Lambda^{*\prime}\Lambda^* = \mathbf{I}_R, \quad \mathbf{C}_{f*}^{(k)} \mathbf{C}_{f*}^{(k)\prime} \quad \text{is diagonal}, \tag{10}$$

where $\mathbf{C}_{f*}^{(k)}$ is defined in the same manner as $\mathbf{C}_f^{(k)}$ in (5) but with replacing F by F^*. Under Assumption A - C, together with (9) and (10), it follows that

$$\begin{aligned} \mathbf{C}_x^{(k)} \mathbf{C}_x^{(k)\prime} &= \Lambda^* \mathbf{C}_{f^*}^{(k)} (\Lambda^{*\prime}\Lambda^*)^{\otimes(k-1)} \mathbf{C}_{f^*}^{(k)\prime} \Lambda^{*\prime} \\ &= \Lambda^* (N^{k-1} \mathbf{C}_{f^*}^{(k)} \mathbf{C}_{f^*}^{(k)\prime}) \Lambda^{*\prime}. \end{aligned} \tag{11}$$

Therefore, Λ^* is $\sqrt{N}$ times the eigenvectors corresponding to the R largest eigenvalues of $\mathbf{C}_x^{(k)} \mathbf{C}_x^{(k)\prime}$.

3. Estimation of the number of non-Gaussian factors

We estimate R based on the eigenvalues of the sample multi-cumulants of x_t. In subsection 3.1, we introduce the generalized eigenvalue ratio (GER) estimator for the case of non-Gaussian factors. In subsection 3.2, we interpret the better finite sample properties of the GER estimator compared to Ahn & Horenstein (2013)'s ER estimator in weak factor models.

3.1. Generalized eigenvalue ratio estimator

We propose to estimate the number of factors R based on the sample higher-order multi-cumulant of the data $\{x_t\}_{t=1}^T$. We define

$$\widetilde{\mu}_{NT,r}^{(k)} \equiv \sigma_r(\widetilde{\mathbf{C}}_x^{(k)}), \tag{12}$$

for $r = 1, 2, \ldots, N$, where the sample k-th order multi-cumulant $\widetilde{\mathbf{C}}_x^{(k)}$ is defined in (6). By maximizing the ratio of the two adjacent singular values of $\widetilde{\mathbf{C}}_x^{(k)}$, we can determine the number of factors. The criterion function is defined as follows:

$$\mathrm{GER}^{(k)}(r) \equiv \frac{\widetilde{\mu}_{NT,r}^{(k)}}{\widetilde{\mu}_{NT,r+1}^{(k)}}, \quad r = 1, 2, \ldots, R_{\max}, \tag{13}$$

where $R_{\max}$ denotes the maximum possible number of factors, $\widetilde{\mu}_{NT,r}^{(k)}$ is defined in equation (12), and $3 \leq k \leq K$, with K being the maximum order considered. We call it the GER estimator because it is an extension of Ahn & Horenstein (2013)'s eigenvalue ratio (ER) estimator. Our proposed estimator for R is the maximizer of $\mathrm{GER}^{(k)}(r)$:

$$\widehat{R}_{\mathrm{GER}}^{(k)} = \max_{1 \leq r \leq R_{\max}} \mathrm{GER}^{(k)}(r). \tag{14}$$

Our main theoretical results for selecting the number of factors are as follows.

Theorem 1. *Suppose that Assumptions A–C hold with $R \geq 1$. If $N^{\alpha}/T \to 0$ as $(N, T) \to \infty$ and $\mathrm{tr}(G_N) = o\big(N^{\frac{k}{k-1}(1-\alpha)} T^{\frac{1}{k-1}} (\log N)^{-\frac{1}{k-1}}\big)$, then,*

(i) *if all factors are skewed ($\phi_j^{(3)} > 0$, $\forall j \in \{1, 2, \ldots, R\}$), we have*

$$\lim_{(N,T)\to\infty} Prob(\widehat{R}_{\mathrm{GER}}^{(3)} = R) = 1,$$

(ii) *if all factors are kurtotic ($\phi_j^{(4)} > 0$, $\forall j \in \{1, 2, \ldots, R\}$), we have*

$$\lim_{(N,T)\to\infty} Prob(\widehat{R}_{\mathrm{GER}}^{(4)} = R) = 1,$$

where $R_{\max} \in (R, N - R - 1]$.

We provide the proof in the Supplementary Appendix. Notice that the GER estimators depend on the quantity $\mathrm{tr}(G_N)$ to detect the number of the non-Gaussian factors. Therefore, we need to study $\mathrm{tr}(G_N)$ to derive the asymptotic properties rather than only assuming $\sigma_1(G_N) < \infty$ as in the existing PCA literatures (Onatski, 2010; Ahn & Horenstein, 2013). Indeed, as also mentioned in Ahn & Horenstein (2013), for many types of high-dimensional datasets, some variables my be almost perfectly correlated with (linear combinations of) others, and thus introduce eigenvalues that are close to zero. This can be further formalized using conditions on the decay of the smallest eigenvalues. Remark 3.1 discusses this condition with some empirical evidence.

Remark 3.1. *Consider the case where G_N has a polynomial rate decaying eigenvalue spectrum given by $\sigma_j(G_N) \leq C_0 j^{-\rho}$ for some $\rho \geq 0$. We refer to ρ as "the decay rate of the spectrum" as in Braun (2006) and Li et al. (2021). Then $\text{tr}(G_N) = O(N^{1-\rho})$ for $0 \leq \rho < 1$, $\text{tr}(G_N) = O(\log N)$ for $\rho = 1$ and $\text{tr}(G_N) = O(1)$ for some $\rho > 1$. In Figure 1, we illustrate that this assumption is realistic. We plot for two empirical datasets the eigenvalues of the sample covariance matrix of error terms after extracting the PCA factors, with the number of PCA factors determined by Ahn & Horenstein (2013)'s Eigenvalue Ratio (ER) estimator. The first one is the FRED-MD dataset of 124 macroeconomic indicators ($N = 124, T = 720$) in the US (McCracken & Ng, 2016), and the second one is the daily returns of S&P 500 component stocks ($N = 449, T = 2263$).[1] The eigenvalues of both datasets are fitted by the polynomial decaying function and represented using blue dashed lines. The parameter ρ in two datasets are 0.753 and 0.718, respectively. The decay rates remain 0.544 and 0.510 after we remove the first ten PCA factors. Therefore, the polynomial decay assumption of the spectrum of G_N is supported by empirical evidence.*

$\sim$ Insert Figure 1 Here $\sim$

The following corollary gives the sample size conditions which are required to ensure the consistency of the GER estimators for a polynomial rate decaying eigenvalue spectrum G_N.

Corollary 1. *Under the conditions in Theorem 1, if G_N has a polynomial rate decaying eigenvalue spectrum such that $\sigma_j(G_N) \leq C_0 j^{-\rho}$ for some $\rho \geq 0$ and positive constant C_0, then we need $\frac{N^{(\alpha-\rho)k} \log N}{N^{1-\rho} T} = o(1)$ and $N^\alpha / T = o(1)$ for $0 \leq \rho < 1$, and $N^\alpha / T = o(1)$ and $N \geq (\log N)^k$ for $\rho \geq 1$ to guarantee the consistency of $\widehat{R}^{(k)}_{\text{GER}}$.*

For $\rho = 0$ such that $\text{tr}(G_N) = O(N)$, e.g. $G_N = \mathbf{I}_N$, we need the time dimension $T \gg N^{k-1} \log N$ to detect the extreme weak factors ($\alpha = 1$). If G_N has a polynomial rate decay spectrum, the requirement of the time dimension T is reduced as $\sigma_j(G_N)$ has a sharper tail.

Remark 3.2. (i) *The parameter K denotes the maximum order considered in the test. As rule of thumb, we recommend $K = 4$ and then consider $k = 3, 4$ in this paper.*

(ii) *We recommend setting $R_{\max}$ similar to Ahn & Horenstein (2013). Please refer to the Supplementary Appendix for more details.*

(iii) *We present a generalization of the ER estimator of the number of factors, as introduced*

[1]The sample period of the FRED-MD data used ranges from January 1959 to December 2018, approximately 60 years of monthly data. The sample period of S&P 500 stocks ranges from July 1, 2010, to June 30, 2019, about 10 years of daily data. We omit several delisted stocks during this period.

by Ahn & Horenstein (2013). They also present a growth ratio estimator, and we give the Generalized Growth Ratio (GGR) estimator in the Supplementary Appendix.

(iv) *In real data analysis, it is difficult to have a priori information on the non-normality of the factors. Hence, we should conduct both* $GER^{(3)}$ *and* $GER^{(4)}$ *and take* $\widehat{R}^* = \max(\widehat{R}^{(3)}_{GER}, \widehat{R}^{(4)}_{GER})$ *to avoid* $\widehat{R}^{(3)}_{GER}$ *and underestimate* R *when the factors are symmetric.*

3.2. Finite sample properties of the GER versus the ER of Ahn & Horenstein (2013)

In this section, we use a stylized setup to show that the GER estimator has better finite sample properties than Ahn & Horenstein (2013)'s ER estimator on detecting weak non-Gaussian factors.

We consider a two-factor model $x_{it} = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + e_{it}$ with $\mathbb{E}(f_t) = \mathbb{E}(e_{it}) = 0$, f_{1t}, f_{2t}, and e_{it} being mutually independent, $\sqrt{N^{1-\alpha}}\lambda_i \sim N(0, \mathbf{I}_R)$, $e_t \sim N(0, G_N)$, where $\sigma_j(G_N) = j^{-\rho}$ with $\rho \geq 0$ for $j = 1, 2, \ldots, N$. We denote the standard deviation of the factors by $sd_j \equiv \sqrt{\mathbb{E}(f_{jt}^2)}$ and the standardized skewness of factors $sk_j^* \equiv sk_j / sd_j^3$ for $j = 1, 2$, where $sk_j \equiv \mathbb{E}(f_{jt}^3)$. We first study the power of the ER estimator based on the eigenvalue decomposition of $\widetilde{\Sigma}_{x,N}$. We can show that

$$\begin{aligned} \sigma_1(\widetilde{\Sigma}_{x,N})/\sigma_2(\widetilde{\Sigma}_{x,N}) &\asymp \frac{sd_1^2}{sd_2^2}, \\ \sigma_2(\widetilde{\Sigma}_{x,N})/\sigma_3(\widetilde{\Sigma}_{x,N}) &\asymp O_p(N^{1-\alpha}) sd_2^2. \end{aligned} \tag{15}$$

When $\alpha = 0$, we can detect two factors because $\sigma_2(\widetilde{\Sigma}_{x,N})/\sigma_3(\widetilde{\Sigma}_{x,N}) \to \infty$ as $N \to \infty$. However, when $0 < \alpha < 1$, for finite N, when f_{2t} is a weaker factor ($sd_2^2 << sd_1^2$), it may hold that $sd_1^2/sd_2^2 \gtrsim O_p(N^{1-\alpha}) sd_2^2$. When $\alpha = 1$, $sd_1^2/sd_2^2 \gtrsim O_p(1) sd_2^2$ can hold even if $N \to \infty$. Therefore, the ER estimator has a low efficiency in the weak factor model ($\alpha > 0$) because it can detect only the strong factor.

Consider now the GER estimator based on the eigenvalue decomposition of the third-order multi-cumulant of x_t. We can show that

$$\begin{aligned} \sigma_1(\widetilde{\mathbf{C}}^{(3)}_{x,N})/\sigma_2(\widetilde{\mathbf{C}}^{(3)}_{x,N}) &\asymp \frac{|sk_1|}{|sk_2|} = \frac{|sk_1^*|}{|sk_2^*|}\frac{sd_1^3}{sd_2^3}, \\ \sigma_2(\widetilde{\mathbf{C}}^{(3)}_{x,N})/\sigma_3(\widetilde{\mathbf{C}}^{(3)}_{x,N}) &\asymp O_p(\min(\sqrt{TN^{1+2\rho-3\alpha}/\log N}, \sqrt{TN^{-\alpha}}))|sk_2^*| sd_2^3. \end{aligned} \tag{16}$$

The singular value ratio of $\widetilde{\mathbf{C}}^{(3)}_{x,N}$ considers the skewness of factors, the first singular value ratio is determined by the skewness ratio of the factors, and the second singular value ratio diverges to infinity as $O_p(\min(\sqrt{TN^{1+2\rho-3\alpha}/\log N}, \sqrt{TN^{-\alpha}})) \to \infty$. When $\alpha = 1$, the second singular value ratio still diverges to infinity if $\min(\frac{N^{2-2\rho}\log N}{T}, \frac{N}{T}) \to 0$ as $(N, T) \to \infty$. For a specific sample size,

if $|sk_2^*|^2/|sk_1^*| \gg O\big(\max(\sqrt{\frac{N^{3-2\alpha}}{T}}, \sqrt{\frac{N^{2-2\rho}\log N}{T}})\big)$ holds, we have

$$\begin{aligned}\sigma_1(\widetilde{\Sigma}_{x,N})/\sigma_2(\widetilde{\Sigma}_{x,N}) &\asymp \sigma_2(\widetilde{\Sigma}_{x,N})/\sigma_3(\widetilde{\Sigma}_{x,N}),\\ \sigma_1(\widetilde{\mathbf{C}}^{(3)}_{x,N})/\sigma_2(\widetilde{\mathbf{C}}^{(3)}_{x,N}) &\ll \sigma_2(\widetilde{\mathbf{C}}^{(3)}_{x,N})/\sigma_3(\widetilde{\mathbf{C}}^{(3)}_{x,N}),\end{aligned} \tag{17}$$

which implies that the GER estimator can detect weaker non-Gaussian factors when the ER estimator has low efficiency. The simulations in the Supplementary Appendix confirm this proposition for the finite sample size. As the eigenvalues of G_N increase, the average maximum eigenvalue ratio of the ER estimator changes from the second one to the first one, which indicates that only the first strong factor is detected. Conversely , the average maximum eigenvalue ratio of the GER estimator remains the second one. As shown in section 8 of the Supplementary Appendix, our proposed GER estimator can also deal with this case efficiently.

4. Estimation of the non-Gaussian factors and loadings

Principal component analysis is the workhorse approach to estimating factors using the eigenvalue decomposition of the sample covariance matrix. We show that this approach can be extended to obtain factors based on an eigenvalue decomposition of $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'}$ $(3 \le k \le K)$. More precisely, the HFA estimate of the factor loadings, denoted by $\widehat{\Lambda}^{(k)}$, are the first R eigenvectors of $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'}$. Given $\widehat{\Lambda}^{(k)}$, the HFA factors are estimated by regression $\widehat{F}^{(k)} = X\widehat{\Lambda}^{(k)}(\widehat{\Lambda}^{(k)\prime}\widehat{\Lambda}^{(k)})^{-1} = X\widehat{\Lambda}^{(k)}/N$. When $k = 2$, the approach nests the PCA-based approach of Bai (2003) as a special case. The estimation of latent factor model parameters based on minimizing the distance of sample and model-based moments is also considered in Jondeau et al. (2018) and Boudt et al. (2020). Their approaches do not lead to an explicit solution. For the approach by Jondeau et al. (2018), the asymptotic properties of factors and factor loadings are unknown. We describe these two approaches in section 5 of the Supplementary Appendix.

In subsection 4.1, we derive the estimates of factors and factor loadings. Subsequently, we establish the convergence rate and the limiting distributions of the estimated factors and factor loadings. In subsection 4.2, we show that the limiting distributions and convergence rates of the HFA estimators obtain an efficiency gain over the PCA estimators in the case of weak non-Gaussian factors. In subsection 4.3, we discuss the feasibility of HFA without the normality of errors. In section 7 of the Supplementary Appendix, we show that the computational cost of the algorithm is moderate when computing the eigenvectors of $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime}$.

4.1. HFA estimators and their asymptotic properties

In this subsection, we give the estimates of the factors and loadings and show they have attractive asymptotic properties. We assume that the number of factors R is known (or estimated consistently). Following (11), we estimate the factor loading matrix Λ up to a rotation matrix by using the following optimization problem:

$$\widehat{\Lambda}^{(k)} = \arg\max_{\Lambda^*} \operatorname{tr}\{\Lambda^{*\prime}(\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime})\Lambda^*\} \tag{18}$$

subject to the constraint $\frac{1}{N}\Lambda^{*\prime}\Lambda^* = \mathbf{I}_R$. Notably, this is the same optimization problem as finding the eigenvectors of the matrix $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime}$. Given $\widehat{\Lambda}^{(k)}$, the factors can be obtained by least squares regression leading to $\widehat{F}^{(k)} = X\widehat{\Lambda}^{(k)}(\widehat{\Lambda}^{(k)\prime}\widehat{\Lambda}^{(k)})^{-1} = X\widehat{\Lambda}^{(k)}/N$.

Remark 4.1. (i) *The constraint of Λ is widespread in covariance-based factor analysis, such as PCA and maximum likelihood analysis. In standard PCA, we maximize* $\operatorname{tr}(\Lambda^{*\prime}\widetilde{\mathbf{C}}_x^{(2)}\Lambda^*)$ *subject to the constraint* $\frac{1}{N}\Lambda^{*\prime}\Lambda^* = \mathbf{I}_R$, *which is equivalent to maximizing* $\operatorname{tr}\{\Lambda^{*\prime}(\widetilde{\mathbf{C}}_x^{(2)}\widetilde{\mathbf{C}}_x^{(2)\prime})\Lambda^*\}$. *Therefore, PCA can be regarded as a special case of HFA when $k = 2$.*

(ii) *For any different $k \geq 3$, we have different loadings $\widehat{\Lambda}^{(k)}$ and different factors $\widehat{F}^{(k)}$. To choose an optimal order k, we suggest using the goodness of fit criterion for the $\mathbf{C}_x^{(k)}$-model in (9). Therefore, we can choose the optimal k by minimizing the partial loss function* $\widehat{k} = \arg\min_k\{\|\widetilde{\mathbf{C}}_x^{(k)} - \widehat{\Lambda}^{(k)}\widehat{\mathbf{C}}_f^{(k)}(\widehat{\Lambda}^{(k)\prime})^{\otimes(k-1)}\|^2/\|\widetilde{\mathbf{C}}_x^{(k)}\|^2\}$, *where $\widehat{\mathbf{C}}_f^{(k)}$ is the k-th order multi-cumulant of $\widehat{F}^{(k)}$. In addition, as shown in the following theorem, the smaller k, the faster the convergence rate of the HFA estimators. Hence, without any prior knowledge, setting up $k = 3$ is a reasonable choice.*

The following theorem provides the rate of convergence of the HFA estimators.

Theorem 2. *Under Assumptions A–C, for any $3 \leq k \leq K$ such that $\phi_j^{(k)} > 0 (\forall\, j \in \{1, 2, \ldots, R\})$, there exists an $R \times R$ invertible matrix $H^{(k)}$ for which*

$$\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(k)} - \Lambda H^{(k)}\| = O_p\Big(\sqrt{\frac{\operatorname{tr}(G_N)^{k-1}\log N}{N^{(1-\alpha)k}T}}\Big) + O_p\Big(\sqrt{\frac{N^\alpha}{T}}\Big),$$
$$\frac{1}{\sqrt{T}}\|\widehat{F}^{(k)} - F(H^{(k)})^{-1}\| = O_p\Big(\frac{1}{\sqrt{N}}\Big) + O_p\Big(\frac{1}{\sqrt{TN^{1-\alpha}}}\Big) + O_p\Big(\sqrt{\frac{\operatorname{tr}(G_N)^{k-1}\log N}{N^{(1-\alpha)k+1}T}}\Big).$$

We provide the proof in the Supplementary Appendix. Theorem 2 shows that the convergence rate of the estimated factors and factor loadings of HFA depend on the factor strength α and $\operatorname{tr}(G_N)$. The following corollary establishes the consistency of the HFA estimates with specific G_N.

Corollary 2. *Under the conditions in Theorem 2, if* $\sigma_j(G_N) \leq C_0 j^{-\rho}$ *for some* $\rho \geq 0$ *and positive constant* C_0 *and* $N \geq (\log N)^k$*, then*

$$\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(k)} - \Lambda H^{(k)}\| = \begin{cases} O_p(\sqrt{\frac{N^{(\alpha-\rho)k}\log N}{TN^{1-\rho}}}) + O_p(\sqrt{\frac{N^\alpha}{T}}) & , \quad 0 \leq \rho < 1, \\ O_p(\sqrt{\frac{N^\alpha}{T}}) & , \quad 1 \leq \rho. \end{cases} \tag{19}$$

$$\frac{1}{\sqrt{T}}\|\widehat{F}^{(k)} - F(H^{(k)})^{-1}\| = \begin{cases} O_p(\sqrt{\frac{N^{(\alpha-\rho)k}\log N}{TN^{2-\rho}}}) + O_p(\frac{1}{\sqrt{N}}) & , \quad 0 \leq \rho < 1, \\ O_p(\frac{1}{\sqrt{N}}) & , \quad 1 \leq \rho. \end{cases} \tag{20}$$

When $\alpha = 0$, namely a strong factor model as in Bai (2003), the convergence rates of $\widehat{F}^{(k)}$ and $\widehat{\Lambda}^{(k)}$ are mainly dominated by $O_p(T^{-1/2})$ and $O_p(N^{-1/2})$. For the case $\alpha > 0$, $\widehat{F}^{(k)}$ and $\widehat{\Lambda}^{(k)}$ are remain consistent under mild conditions with respect to sample size (N, T), factor strength α, and decay rate ρ. Notably, the rotation matrix $H^{(k)}$ is different from the rotation matrix in Bai (2003). $H^{(k)}$ contains the higher-order multi-cumulants of the non-Gaussian factors. Moreover, $H^{(k)}$ depends on the parameter k $(3 \leq k \leq K)$, the order of the multi-cumulant we used in the estimation.

To derive the limit distributions of $\widehat{f}_t^{(k)}$ and $\widehat{\lambda}_i^{(k)}$, we require the following additional assumptions:

ASSUMPTION D: Central limit theorem

(i) Let k $(3 \leq k \leq K)$ be such that it satisfies Assumption A(ii). Subsequently, for each i, as $T \to \infty$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\zeta_t^{(k)} e_{it} \xrightarrow{d} \mathcal{N}(0, \Theta_i^{(k)}),$$

where $\zeta_t^{(k)} \in \mathbb{R}^{R^{k-1}}$ is the t-th row of the matrix $\bar{\mathcal{H}}_f^{(k)}$ and $\Theta_i^{(k)} = p\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}(\zeta_t^{(k)}\zeta_s^{(k)\prime}e_{it}e_{is})$. The matrix $\bar{\mathcal{H}}_f^{(k)} \in \mathbb{R}^{T\times R^{k-1}}$ satisfies $T^{-1}F'\bar{\mathcal{H}}_f^{(k)} = \widetilde{\mathbf{C}}_f^{(k)}$. The explicit form of matrix $\bar{\mathcal{H}}_f^{(k)}$ $(k = 3, 4)$ is given in Remark 4.2.

(ii) For each t, as $N \to \infty$,

$$\frac{1}{\sqrt{N^{1-\alpha}}}\sum_{i=1}^{N}\lambda_i e_{it} \xrightarrow{d} \mathcal{N}(0, \Phi_t),$$

where $\Phi_t = \lim_{N\to\infty}\frac{1}{N^{1-\alpha}}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}(\lambda_i\lambda_j' e_{it}e_{jt})$ and $\alpha \in [0, 1]$.

Assumption D(i) requires the asymptotic normality of the cross product between $\zeta_t^{(k)}$ and e_{it}, which depends on the order k. When $k = 2$, Remark 4.2 implies that $\bar{\mathcal{H}}_f^{(2)} = F$. Hence, this

assumption is equivalent to Bai (2003)'s Assumption F. Notice that $\Theta_i^{(k)}$ is an $R^{k-1} \times R^{k-1}$ matrix, and Φ_t is an $R \times R$ matrix. Assumption D(ii) requires the asymptotic normality of the cross product between λ_i and e_{it}, and the CLT for the factor loadings when $\alpha = 0$ has the same form as in Bai (2003)'s Assumption F. If $\alpha > 0$, we allow for a slower convergence rate than $\sqrt{N}$. This is because $\|\lambda_i\| = O(N^{-\frac{\alpha}{2}})$ by Assumptions B(ii). Subsequently, the following result holds:

Theorem 3. *Under Assumptions A–D, for any* $3 \leq k \leq K$ *such that* $\phi_j^{(k)} > 0 (\forall\, j \in \{1, 2, \ldots, R\})$, *then*

(i) *if* $N^\alpha / T \to 0$ *and* $\mathrm{tr}(G_N) = o(N^{\frac{k}{k-1} - \alpha} (\log N)^{-\frac{1}{k-1}})$, *we have*

$$\sqrt{TN^{-\alpha}} \Big(\widehat{\lambda}_i^{(k)} - H^{(k)'} \lambda_i \Big) \xrightarrow{d} \mathcal{N} \Big(0, (D^{(k)})^{-1} Q^{(k)'} \mathbf{C}_f^{(k)} (\Sigma_\Lambda)^{\otimes k-1} \Theta_i^{(k)} (\Sigma_\Lambda)^{\otimes k-1} \mathbf{C}_f^{(k)'} Q^{(k)} (D^{(k)})^{-1} \Big),$$

(ii) *if* $N^\alpha / T \to 0$ *and* $\mathrm{tr}(G_N) = o(T^{\frac{1}{k-1}} N^{\frac{k}{k-1}(1-\alpha)} (\log N)^{-\frac{1}{k-1}})$, *we have*

$$\sqrt{N} \Big(\widehat{f}_t^{(k)} - (H^{(k)})^{-1} f_t \Big) \xrightarrow{d} \mathcal{N} \Big(0, (Q^{(k)'})^{-1} \Phi_t (Q^{(k)})^{-1} \Big),$$

where $Q^{(k)} = (\Psi^{(k)})^{-1/2} \Gamma^{(k)} (D^{(k)})^{1/2}$, $\Psi^{(k)} = \mathbf{C}_f^{(k)} (\Sigma_\Lambda)^{\otimes (k-1)} \mathbf{C}_f^{(k)'}$, $D^{(k)}$ *is the diagonal eigenvalue matrix of* $(\Psi^{(k)})^{1/2} \Sigma_\Lambda (\Psi^{(k)})^{1/2}$, $\Sigma_\Lambda = \lim_{N \to \infty} \Lambda' \Lambda / N^{1-\alpha}$, $\Gamma^{(k)}$ *is the corresponding eigenvector matrix such that* $\Gamma^{(k)'} \Gamma^{(k)} = \mathbf{I}_R$.

We provide the proof in the Supplementary Appendix. The limit distribution of the HFA estimators has a more generalized form and depends on the order k. Theorem 3 has two noteworthy points. First, the asymptotic normality of the estimated factors $\widehat{f}_t^{(k)}$ holds for all $\alpha \in [0, 1]$ if time dimension T sufficiently larger than cross section dimension N. However, the estimated factor loadings $\widehat{\lambda}_i^{(k)}$ share the asymptotic normality only when $\mathrm{tr}(G_N)$ and factor strength α satisfies specific conditions, e.g., if $\mathrm{tr}(G_N) = O(N)$, then $\widehat{\lambda}_i^{(k)}$ is asymptotically normal for $\alpha < \frac{1}{k-1}$ if we ignore the logarithmic term. Second, if G_N has a polynomial decaying spectrum such that $\mathrm{tr}(G_N) = o(N)$, the demand of T to guarantee the normality of estimated factors is smaller. Additionally, the limit distribution of HFA estimators shares the same form as Bai & Ng (2013)'s PC1 when $k = 2$ and $\alpha = 0$.

Remark 4.2. *The matrices* $\bar{\mathcal{H}}_f^{(k)}$ *and* $\bar{\mathcal{H}}_e^{(k)}$ *share the same form. Thus, we use* $z_t \in \mathbb{R}^{Q \times 1}$ *to derive the matrix* $\bar{\mathcal{H}}_z^{(k)}$ *that satisfies* $T^{-1} Z' \bar{\mathcal{H}}_z^{(k)} = \widetilde{\mathbf{C}}_z^{(k)}$ *for* $k = 3, 4$. *When* $k = 3$, *which is implied by Lemma 2, the matrix* $\bar{\mathcal{H}}_z^{(3)}$ *can be written as*

$$\bar{\mathcal{H}}_z^{(3)} = \big(z_1 \circ z_1, z_1 \circ z_2, \ldots, z_1 \circ z_Q | \ldots | z_Q \circ z_1, z_Q \circ z_2, \ldots, z_Q \circ z_Q \big) \in \mathbb{R}^{T \times Q^2},$$

where $z_i = (z_{i1}, z_{i2}, \ldots, z_{iT})$ *and* $\circ$ *denote the Hadamard product. When* $k = 4$*, according to Lemma 2, the matrix* $\bar{\mathcal{H}}_z^{(4)}$ *can be written as*

$$\begin{aligned}\bar{\mathcal{H}}_z^{(4)} = \big(&z_1 \circ z_1 \circ z_1 - \widetilde{m}_{z,11} z_1 - \widetilde{m}_{z,11} z_1 - \widetilde{m}_{z,11} z_1, \ldots, \\ &z_{i_2} \circ z_{i_3} \circ z_{i_4} - \widetilde{m}_{z,i_3 i_4} z_{i_2} - \widetilde{m}_{z,i_2 i_4} z_{i_3} - \widetilde{m}_{z,i_2 i_3} z_{i_4}, \ldots, \\ &z_Q \circ z_Q \circ z_Q - \widetilde{m}_{z,QQ} z_Q - \widetilde{m}_{z,QQ} z_Q - \widetilde{m}_{z,QQ} z_Q\big) \in \mathbb{R}^{T \times Q^3},\end{aligned} \tag{21}$$

where $\widetilde{m}_{z,i_2 i_3} = \frac{1}{T}\sum_{t=1}^{T} z_{i_2 t} z_{i_3 t}$ *is the sample covariance of* z_{i_2} *and* z_{i_3}*. Note that* $\{z_{i_2} \circ z_{i_3} \circ z_{i_4} - \widetilde{m}_{z,i_3 i_4} z_{i_2} - \widetilde{m}_{z,i_2 i_4} z_{i_3} - \widetilde{m}_{z,i_2 i_3} z_{i_4}\}$ *is the* $\{i_2 + (i_3 - 1)Q + (i_4 - 1)Q^2\}$*-th column of* $\bar{\mathcal{H}}_z^{(4)}$*.*

Remark 4.3. *Let us rotate the underlying factors and loadings to satisfy the normalization conditions such that* $\frac{1}{N}\Lambda'\Lambda = \mathbf{I}_R$ *and* $\widetilde{\mathbf{C}}_f^{(k)} \widetilde{\mathbf{C}}_f^{(k)'}$ *is diagonal. In the Supplementary Appendix, we show that the rotation matrix* $H^{(k)} \to \mathbf{I}_R$*. Thus, the results of Theorem 2 can be simplified as*

$$\begin{aligned}\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(k)} - \Lambda\| &= O_p\Big(\sqrt{\frac{\mathrm{tr}(G_N)^{k-1}\log N}{N^{(1-\alpha)k}T}}\Big) + O_p\Big(\sqrt{\frac{N^\alpha}{T}}\Big), \\ \frac{1}{\sqrt{T}}\|\widehat{F}^{(k)} - F\| &= O_p\Big(\frac{1}{\sqrt{N}}\Big) + O_p\Big(\frac{1}{\sqrt{TN^{1-\alpha}}}\Big) + O_p\Big(\sqrt{\frac{\mathrm{tr}(G_N)^{k-1}\log N}{N^{(1-\alpha)k+1}T}}\Big).\end{aligned} \tag{22}$$

Furthermore, the results of Theorem 3 can be simplified as

$$\begin{aligned}\sqrt{TN^{-\alpha}}\Big(\widehat{\lambda}_i^{(k)} - \lambda_i\Big) &\xrightarrow{d} \mathcal{N}\Big(0, (D^{(k)})^{-1}\mathbf{C}_f^{(k)}\Theta_i^{(k)}\mathbf{C}_f^{(k)'}(D^{(k)})^{-1}\Big), \\ \sqrt{N}\Big(\widehat{f}_t^{(k)} - f_t\Big) &\xrightarrow{d} \mathcal{N}\Big(0, \Phi_t\Big),\end{aligned} \tag{23}$$

where $D^{(k)}$ *is a diagonal matrix and* $\{D^{(k)}\}_{ii} = \sigma_i^2(\mathbf{C}_f^{(k)})$ *for* $i = 1, 2, \ldots, R$*.*

4.2. Efficiency gain of HFA estimators

The HFA estimators are more efficient than the PCA estimators in the case of the weak factor model in estimating non-Gaussian factors ($\alpha > 0$ and $\mathrm{rank}(\mathbf{C}_f^{(k)}) = R$). We illustrate the efficiency gain of HFA estimators compared to PCA estimators in terms of the convergence rate and asymptotic variance. As both the HFA and PCA estimators are the rotation of true factors and factor loadings, we use the normalization $\frac{1}{N}\Lambda'\Lambda = \mathbf{I}_R$ and $\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)'}$ being diagonal to compare them without rotation. Denote the PCA estimators as $(\widehat{f}_t^{PCA}, \widehat{\lambda}_i^{PCA})$. Assumptions A – C imply

the following convergence rate for the PCA estimators:

$$\begin{aligned}\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{PCA}-\Lambda\| &= O_p(\frac{1}{\sqrt{TN^{-\alpha}}})+O_p(\frac{1}{N^{1-\alpha}}),\\ \frac{1}{\sqrt{T}}\|\widehat{F}^{PCA}-F\| &= O_p(\frac{1}{\sqrt{N}})+O_p(\frac{1}{N^{1-\alpha}})+O_p(\frac{1}{\sqrt{TN^{-\alpha}}}).\end{aligned} \tag{24}$$

The convergence rate of the PCA estimator shares the same form as that of the HFA estimator when $\alpha = 0$, which implies that the PCA and HFA estimators have the same convergence rate in the classical strong factor model. However, when $\alpha > 0$, the HFA estimators converge faster than the PCA estimators if $\mathrm{tr}(G_N) = o\big(\sqrt{T}N^{1-\alpha}(\log N)^{-\frac{1}{2}}\big)$. In particular, in the case of $\alpha = 1$, the PCA estimators are inconsistent. Nevertheless, the HFA estimators remain consistent as shown in Theorem 2.

In the case of asymptotic distribution, the PCA estimators (Bai & Ng, 2013) have the form as follows when $(N,T) \to \infty$, $\sqrt{N}/T \to 0$, and $\sqrt{T}/N \to 0$ in a classical strong factor model ($\alpha = 0$):

$$\begin{aligned}\sqrt{T}\big(\widehat{\lambda}_i^{PCA}-\lambda_i\big) &\xrightarrow{d} \mathcal{N}\big(0,(D)^{-1}\Theta_i(D)^{-1}\big),\\ \sqrt{N}\big(\widehat{f}_t^{PCA}-f_t\big) &\xrightarrow{d} \mathcal{N}\big(0,\Phi_t\big),\end{aligned} \tag{25}$$

where $D = p\lim_{T\to\infty} F'F/T$, $\Theta_i = p\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}(f_t f_s' e_{it}e_{is})$, and $\Phi_t = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}(\lambda_i\lambda_j' e_{it}e_{jt})$. We already showed that the HFA estimators in strong factor model ($\alpha = 0$) satisfies:

$$\begin{aligned}\sqrt{T}\big(\widehat{\lambda}_i^{(k)}-\lambda_i\big) &\xrightarrow{d} \mathcal{N}\big(0,(D^{(k)})^{-1}\mathbf{C}_f^{(k)}\Theta_i^{(k)}\mathbf{C}_f^{(k)'}(D^{(k)})^{-1}\big),\\ \sqrt{N}\big(\widehat{f}_t^{(k)}-f_t\big) &\xrightarrow{d} \mathcal{N}\big(0,\Phi_t\big),\end{aligned} \tag{26}$$

where $\{D^{(k)}\}_{ii} = \sigma_i^2(\mathbf{C}_f^{(k)})$ and $\Theta_i^{(k)} = p\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}(\zeta_t^{(k)}\zeta_s^{(k)'}e_{it}e_{is})$.

We mainly compare the asymptotic variance of the estimated factors and factor loadings of PCA and HFA. By Assumption D(ii) and the definition of Φ_t, we have $\widehat{f}_t^{(k)}$ and $\widehat{f}_t^{PCA}$ converge to the same distribution. However, as shown in Theorem 2, $\widehat{f}_t^{(k)}$ converges faster to this distribution than $\widehat{f}_t^{PCA}$ when $0 < \alpha < 1$. Conversely, the variances of $\widehat{\lambda}_i^{(k)}$ and $\widehat{\lambda}_i^{PCA}$ share the sandwich structure, whereas $\Theta_i^{(k)}$ and Θ_i have a different form. Generally, if we set $k = 2$, we have $\Theta_i^{(k)} = \Theta_i$. Therefore, the PCA estimators can be regarded as a special case of the HFA estimator when $k = 2$. The structure of $\Theta_i^{(k)}$ mainly depends on the random variable $\bar{\mathcal{H}}_f^{(k)'}e_i$. When e_i is heteroscedastic and time-series correlated, it is difficult to give an explicit form of $\Theta_i^{(k)}$ and compare it with Θ_i. To better understand the variance difference between $\widehat{\lambda}_i^{(k)}$ and $\widehat{\lambda}_i^{PCA}$, we consider independently

and identically distributed e_{it} with $\mathbb{E}(e_{it}) = 0$ and $\mathbb{E}(e_{it}^2) = sd_e^2$ $(i = 1, 2, \ldots, N)$. Similarly, denote $sd_j^2 \equiv \mathbb{E}(f_{jt}^2)$, $sk_j^* \equiv \mathbb{E}(f_{jt}^3)$, and $kt_j \equiv \mathbb{E}(f_{jt}^4)$ $(j = 1, 2, \ldots, R)$. The order $k = 3$ is used for HFA estimators. Therefore, the variance of $\sqrt{T}(\widehat{\lambda}_{ij}^{PCA} - \lambda_{ij})$ is

$$(D^{-1}\Theta_i D^{-1})_{jj} = (sd_j^2)^{-1}(sd_j^2 sd_e^2)(sd_j^2)^{-1} = sd_e^2/sd_j^2, \tag{27}$$

which indicates that the variance of j-th factor loading is equal to the inverse of Signal-Noise Ratio. For HFA estimators, the variance of $\sqrt{T}(\widehat{\lambda}_{ij}^{(3)} - \lambda_{ij})$ is

$$\begin{aligned}((D^{(3)})^{-1}\mathbf{C}_f^{(3)}\Theta_i^{(3)}\mathbf{C}_f^{(3)'}(D^{(3)})^{-1})_{jj} &= (sk_j)^{-2}(sk_j kt_j sd_e^2 sk_j)(sk_j)^{-2} \\ &= kt_j sd_e^2/(sk_j)^2.\end{aligned} \tag{28}$$

Notice that the Hölder inequality $\mathbb{E}(f_j^2)\mathbb{E}(f_j^4) \geq \mathbb{E}(f_j^3)^2$ always holds. Thus, we have $\text{Var}(\widehat{\lambda}_{ij}^{PCA}) \leq \text{Var}(\widehat{\lambda}_{ij}^{(k)})$ for $j = 1, 2, \ldots, R$. Hence, the estimated factor loadings of HFA have a larger asymptotic variance than the PCA estimators.

To summarize, we find that the HFA factors $\widehat{f}_t^{(k)}$ are consistent irrespective of the value of α. The PCA $\widehat{f}_t^{PCA}$ and HFA $\widehat{f}_t^{(k)}$ factors converge to the same distribution, but the HFA factors converge much faster than the PCA factors when $\alpha > 0$ under mild conditions. Similarly, $\widehat{\lambda}_i^{(k)}$ always guarantees consistency and converges faster than $\widehat{\lambda}_i^{PCA}$ when $\alpha > 0$. Although the asymptotic variance of $\widehat{\lambda}_i^{(k)}$ is larger than $\widehat{\lambda}_i^{PCA}$, we find in the simulation analysis in subsection 6.3 that $\widehat{\lambda}_i^{(k)}$ has better finite sample accuracy as $\alpha > 0$ because the advantage of the faster convergence rate dominates.

4.3. The case of non-Gaussian errors

The results of Theorem 1-3 are all based on the sufficient assumption that the error terms are normally distributed. In this subsection, we discuss the feasibility of HFA in identifying non-Gaussian factors without the normality of u_{it} in Assumption C.

Following the same argument in subsection 2.2, we can derive $\mathbf{C}_x^{(k)} = \Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)} + \mathbf{C}_e^{(k)}$ for $3 \leq k \leq K$ under Assumption A and B. By Weyl's inequality, we have

$$\sigma_j\big(\mathbf{C}_x^{(k)}\big) \asymp \sigma_j\big(\Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)}\big) + \sigma_1\big(\mathbf{C}_e^{(k)}\big), \quad j = 1, \ldots, N. \tag{29}$$

Notice that $\text{rank}(\Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)}) = R$. It follows that

$$\sigma_j\big(\mathbf{C}_x^{(k)}\big) \asymp \begin{cases} \sigma_j\big(\Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)}\big) + \sigma_1\big(\mathbf{C}_e^{(k)}\big), & j = 1, \ldots, R; \\ \sigma_1\big(\mathbf{C}_e^{(k)}\big), & j = R+1, \ldots, N. \end{cases} \tag{30}$$

Under Assumption A and B, we have $\sigma_j\big(N^{(\alpha-1)k/2}\Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)}\big) = O(1)$. If we have

$$\sigma_R\big(\Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)}\big) \gg \sigma_1\big(\mathbf{C}_e^{(k)}\big), \tag{31}$$

then

$$\sigma_j\big(N^{(\alpha-1)k/2}\mathbf{C}_x^{(k)}\big) \asymp \begin{cases} \sigma_j\big(N^{(\alpha-1)k/2}\Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)}\big) = O(1) & , j = 1,\ldots,R; \\ o(1) & , j = R+1,\ldots,N. \end{cases} \tag{32}$$

In other words, when the eigenvalues of the k-th order multi-cumulant of the observable x_t are dominated by the eigenvalues of its non-Gaussian component, then the low rank structure $\Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)}$ can be detected by $\mathbf{C}_x^{(k)}$. Therefore, (31) is the sufficient condition to detect the factors and loadings by the the k-th order multi-cumulant $\mathbf{C}_x^{(k)}$ without the normality assumption of the error terms.

In the following proposition, we propose mild conditions under which the sufficient condition (31) is still satisfied when the errors are non-normal. The condition depends on the factor strength α, $\sigma_1(G_N)$, $\mathrm{tr}(G_N)$, the number of nonzero elements in G_N and the k-th order cumulant of u_{it}.

Proposition 1. *Let $\{u_{it}\}_{t=1}^T$ be a strong mixing sequence such that $\mathbb{E}(u_{it}) = 0$ and $\mathbb{E}(u_{it}^{2K}) < \infty$. Let $g_1^* \geq \ldots \geq g_N^*$ be the eigenvalues of G_N and $G_N^* = \mathrm{diag}(g_1^*,\ldots,g_N^*)$. Let $L = (l_1,\ldots,l_N)$ be the eigenvector matrix of G_N such that $G_N = LG_N^*L'$. Let $\mathcal{G}_i = \sum_{j=1}^N \mathbf{1}\{l_{ji} \neq 0\}$ be the number of non-zero elements of l_i. Assume $\kappa_{u,i}^{(k)} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \kappa_{u,i,t}^{(k)}$ exists and $\sqrt{\mathcal{G}_i}|l_{ji}| = O(1)$ if $l_{ji} \neq 0$. Let $\mathcal{G} = \min_i \mathcal{G}_i$ and $\kappa_u^{(k)} = \max_i |\kappa_{u,i}^{(k)}|$. Then under Assumption A and B, we have $\sigma_R(\Lambda\mathbf{C}_f^{(k)}\Lambda'^{\otimes(k-1)}) \gg \sigma_1(\mathbf{C}_e^{(k)})$, provided that*

$$N^{\frac{(\alpha-1)k}{2}}\sigma_1(G_N)[\mathcal{G}^{-1}\mathrm{tr}(G_N)]^{\frac{k}{2}-1}\kappa_u^{(k)} \quad = \quad o(1) \tag{33}$$

holds for any eigenvector matrix L belonging to the eigenvector space of G_N.

We provide the proof in the Supplementary Appendix. Proposition 1 nests the condition of the Gaussian error term if we take $\kappa_u^{(k)} = 0$. If $\kappa_u^{(k)} \neq 0$, we need smaller $\mathrm{tr}(G_N)$ and larger $\mathcal{G}$ to detect weak factors. For example, $\mathrm{tr}(G_N)/\mathcal{G} = O(1)$, $\sigma_1(G_N) < \infty$ and $\mathcal{G} \to \infty$ as $N \to \infty$ is sufficient to detect the factors with $\alpha \in [0,1)$. The worst case for HFA is $\kappa_u^{(k)} \neq 0$, $\mathcal{G} = 1$, $\mathrm{tr}(G_N) = O(N)$, e.g. $G_N = \mathbf{I}_N$, then HFA can only detect the factors with $\alpha \in [0, 2/k)$.

5. Forecasts in factor-augmented regressions

The important use of factors to achieve dimension reduction has been found to be empirically useful in analyzing macroeconomic time series. Adding factors to a forecasting model is being used

by an increasing number of researchers. See for example, Stock & Watson (2002), Bai & Ng (2006) and McCracken & Ng (2016). In this section, we establish the rate of convergence and limiting distribution for the estimated parameters of the factor-augmented regression with the HFA factors. Subsequently, we derive similar results for the predicted conditional mean and the forecasting error. For predictive inference, we also give the estimates of the variance of forecasts.

Suppose information is available on a large number of predictors $x_{it}(i = 1, 2, \ldots, N; t = 1, 2, \ldots, T)$ and other small set of observable variables W_t (such as lags of y_t). Consider

$$y_{t+h} = \beta' f_t + \gamma' W_t + \epsilon_{t+h}, \tag{34}$$

where $h \geq 0$ is the lag time between the dependent variable y_t and available information (f_t, W_t). The vector f_t is unobservable and comes from a large panel of data x_{it}. We refer to a weak factor model $x_t = \Lambda f_t + e_t$ with $\|\Lambda'\Lambda\| \asymp N^{1-\alpha}$ and $\alpha \in [0, 1]$. If f_t is observable and assuming the mean of ϵ_t conditional on past information is zero, the optimal prediction of y_{T+h} on time T is the conditional mean and is given by

$$y_{T+h|T} = \mathbb{E}[y_{T+h}|z_T, z_{T-1}, \ldots] = \beta' f_t + \gamma' W_t \equiv \delta' z_T, \tag{35}$$

where $z_t = (f_t', W_t)'$. However, this prediction is not feasible because δ and f_t are all unobserved. Replacing f_t and δ by their estimates, we obtain a feasible prediction as

$$\widehat{y}_{T+h|T} = \widehat{\beta}' \widehat{f}_T^{(k)} + \widehat{\gamma}' W_T = \widehat{\delta}' \widehat{z}_T^{(k)}, \tag{36}$$

where $\widehat{z}_t^{(k)} = (\sqrt{N^\alpha} \widehat{f}_t^{(k)'}, W_t)'$, and $\widehat{f}_t^{(k)}$ are the HFA estimates of the factor model (2) with normalization $\widehat{\Lambda}^{(k)'} \widehat{\Lambda}^{(k)}/N = \mathbf{I}_R$. $\widehat{\beta}$ and $\widehat{\gamma}$ are the least squares estimates obtained from a regression of y_{t+h} on $\widehat{f}_t^{(k)}$ and W_t, $t = 1, 2, \ldots, T - h$. Bai & Ng (2006) study the statistical properties of $\widehat{y}_{T+h|T}$ and the estimated parameters $\widehat{\delta}$ in a strong factor model ($\alpha = 0$). Moreover, they show the asymptotic normality of $\widehat{\delta}$ based on PCA factors if $\sqrt{T}/N \to 0$, and the prediction confidence interval of $\widehat{y}_{T+h|T}$ is further given. Nevertheless, the asymptotic properties of the factor-augmented regression based on weak factors have not been studied yet. To evaluate the uncertainty of a diffusion index forecast, we need the limiting distributions of $\widehat{\delta}$ and $\widehat{y}_{T+h|T}$. In section 2 of the Supplementary Appendix, we set up these results based on the k-th order HFA factors $\widehat{f}_t^{(k)}$. As we will see, by using the HFA factors, the diffusion index parameters and forecasts ensure consistency and asymptotic normality even when the factor strength is considerably weak ($\alpha \to 1$).

6. Simulation studies

This section reports the results from several Monte Carlo simulations regarding the performance of our proposed HFA methodology in finite samples. We define the simulation set up in subsection 6.1. In subsections 6.2 and 6.3, we consider a base scenario in which all factors are non-Gaussian because we are interested in the improvement in the case of non-Gaussian factors. In subsection 6.4, we study the robustness of our findings in sensitivity analysis of the HFA estimators. Additional robustness checks are presented in section 8 of the Supplementary Appendix.

6.1. Simulation set up

We consider two types of data generation processes (DGP) for the simulations. DGP1 uses the following factor model:

$$\begin{aligned} x_{it} &= \sum_{r=1}^{R} \lambda_{ir} f_{rt} + e_{it}, \quad e_t = G_N^{1/2} u_t, \\ f_{jt} &= d_j f_{jt-1} + v_{jt} \quad , \quad v_{jt} \backsim i.i.d. SGT(0,1,\eta_j,p_j,q_j), \\ u_{it} &= \xi u_{it-1} + u_{it}^* \quad , \quad u_{it}^* \backsim i.i.d. \mathcal{N}(0,1), \\ \lambda_{ij} &\backsim i.i.d. \mathcal{N}(0, N^{\alpha-1}), \end{aligned} \tag{37}$$

where $\sigma_n(G_N) = n^{-0.544}$, which is calibrated by the FRED-MD dataset after removing the first ten PCA factors. DGP1 uses the eigenvalues of G_N to control the signal-noise ratio. DGP2 follows the factor model used in Bai & Ng (2002) and Ahn & Horenstein (2013):

$$\begin{aligned} x_{it} &= \sum_{r=1}^{R} \lambda_{ir} f_{rt} + \sqrt{\theta_i} e_{it}, \quad e_{it} = \sqrt{\frac{1-\xi^2}{1+2J\beta^2}} u_{it}^*, \\ u_{it} &= \xi u_{it-1} + u_{it}^* + \sum\nolimits_{h=\max(i-J,1)}^{i-1} \beta u_{ht}^* + \sum\nolimits_{h=i+1}^{\min(i+J,N)} \beta u_{ht}^*, \\ f_{jt} &= d_j f_{jt-1} + v_{jt}, \quad v_{jt} \backsim i.i.d.\, SGT(0,1,\eta_j,p_j,q_j), \\ u_{it}^* &\backsim i.i.d. \mathcal{N}(0,1), \quad \lambda_{ij} \backsim i.i.d. \mathcal{N}(0,1). \end{aligned} \tag{38}$$

The parameter ξ controls the magnitude of the serial and cross-sectional correlation, which is governed by two parameters — β and J — which specify the magnitude of the cross-sectional correlation and number of correlated cross-sectional units, respectively. The parameter θ_i is the variance of each idiosyncratic error.

We use the skewed generalized error (SGT) distribution to describe the non-normality of the factors and idiosyncratic errors, where the distribution $SGT(\mu,\sigma,\eta,p,q)$ is a univariate 5-parameter

distribution introduced by Theodossiou (1998) and known for its extreme flexibility. The SGT distribution has the probability density function

$$f_{SGT}(x;\mu,\sigma,\eta,p,q) = \begin{cases} \zeta\left[(1+\frac{p}{q-2})v^{-p}(1-\eta)^{-p}|\frac{x-\mu}{\sigma}|^p\right]^{-\frac{q+1}{p}}, & x<\mu, \\ \zeta\left[(1+\frac{p}{q-2})v^{-p}(1+\eta)^{-p}|\frac{x-\mu}{\sigma}|^p\right]^{-\frac{q+1}{p}}, & x\geq\mu. \end{cases} \tag{39}$$

where ζ and v are normalizing constants ensuring that $f_{SGT}(\cdot)$ is a proper probability density function. The parameter μ is the mean, and σ is the standard deviation of the distribution. The parameter η determines the skewness; p and q determine the kurtosis. The k-th moment exists when $pq > k$.

As stated earlier, we set the number of factors to be 3 ($R = 3$). The distribution parameters of the non-Gaussian factors are considered as $\eta_j = 0.5, p_j = 1, q_j = \infty$ such that the factors have unit variance, 1.244 skewness and 4.920 excess-kurtosis. For DGP1, we merely need to change α to control the strong or weak factor model. For DGP2, we fix the correlation structures of the error terms as $\xi = 0.2, \beta = 0.2$, $J = [N/10]$, and $\theta_i \backsim U[1,\theta]$ and then change θ to control the strong or weak factor model.

6.2. Finite sample properties of the GER estimator

In this subsection, we evaluate the finite sample properties of the GER estimator. The performances of the two estimators are compared with covariance-based estimators, for example, the PC3 estimator of Bai & Ng (2002), the ON estimator of Onatski (2010), and the ER and GR estimators of Ahn & Horenstein (2013). We also compare with Jondeau et al. (2018)'s JJR method, which is based on the higher-order moment (see in section 6 of the Supplementary Appendix). We focus on how the finite sample properties of those estimators are affected by the parameters α (in DGP1) and θ (in DGP2) — the strong or weak factor model. When α or θ is small, the influential power of all factors is strong, and the factors are gradually weakened as α or θ increases. We set the maximum number of factors $R_{\max} = 10$ and consider $T \in \{300, 500, 1000\}$ and $N \in \{100, 300\}$.

Figure 2 reports the finite sample performances of the above estimators when $\alpha \in [0, 1]$ and $\theta \in [1, 10]$. The GER estimator is based on the third-order multi-cumulants; for each estimator, we run 500 replications. The results in Figure 2 show that the GER estimator has good finite sample properties in both strong or weak factor models, particularly in the weak factor case in which all covariance-based methods have low efficiency. Additionally, the JJR method shows inefficient in both DGP1 and DGP2 even for small α and θ.

∼ Insert Figure 2 Here ∼

6.3. Finite sample properties of HFA factors and factor loadings

In this subsection, we evaluate the finite sample properties of the PCA and HFA estimators of the factors and factor loadings in the previous DGPs and assume that the number of factors $R = 3$ are known. We denote the PCA estimators by $\widehat{F}^{PCA}$ and $\widehat{\Lambda}^{PCA}$ and the HFA estimators by $\widehat{F}^{HFA}$ and $\widehat{\Lambda}^{HFA}$; the latter are obtained using the third-order multi-cumulant ($k = 3$). As a measure of goodness-of-fit, we use the trace ratio (TR) to evaluate how close the estimated values $\widehat{\Lambda}$ and $\widehat{F}$ are to their true values. Taking $\widehat{F}$ as an example, the TR is defined as

$$\text{TR}(\widehat{F}, F) = \frac{\text{tr}((F'\widehat{F})(\widehat{F}'\widehat{F})^{-1}(\widehat{F}'F))}{\text{tr}(F'F)}. \tag{40}$$

The measure is a generalized squared correlation coefficient in multivariate analysis and is invariant to rotation. It is widely used as a measure of goodness-of-fit in factor analysis; see, for example, Bai et al. (2012) and Bai & Li (2016). Figure 3 reports the average TR based on 500 repetitions for each (N, T) combination under $\alpha \in [0, 1]$ or $\theta \in [1, 10]$. We can observe the following points. First, for each sample size (N, T), the PCA estimators have a lower efficiency than the HFA estimators when α or θ is large, which is expected because the HFA estimators, as shown in Theorem 2, converge faster than the PCA estimators. Second, when α or θ is small, $\widehat{\Lambda}^{PCA}$ show a smaller efficiency gain than $\widehat{\Lambda}^{HFA}$ because $\widehat{\Lambda}^{HFA}$ have larger asymptotic variance than $\widehat{\Lambda}^{PCA}$. Conversely, $\widehat{F}^{HFA}$ always outperforms than $\widehat{F}^{PCA}$ regardless of what α or θ is. Third, the finite sample properties of $\widehat{\Lambda}$ and $\widehat{F}$ are dominated by T and N, respectively. Overall, the simulations confirm our theorems and also the better performance of the HFA estimators, particularly in weak factor cases.

∼ Insert Figure 3 Here ∼

6.4. Sensitivity analysis

In the above simulations, we assume that the factor distributions are highly non-Gaussian and that the error distributions are Gaussian to study the finite sample properties of the HFA estimators. In this sensitivity analysis, we study how the finite sample properties change when only "mild non-Gaussianity" exists in factors and when the idiosyncratic errors are non-Gaussian? We also analyze the sensitivity of the HFA estimation performance to the decay rate of the spectrum of G_N.

First, we evaluate the sensitivity of the HFA estimators concerning the strength of the higher-order cumulant of factors. We follow the three-factor model in (37) and change the skewness of all factors from zero to two (for SGT distribution, we change η_j from 0 to 0.98) and α from zero to one in DGP1. The sample size is $(N, T) = (300, 500)$. For each possible combination of η_j and α, we compute the frequency of the correct estimation of the GER3 estimator, the average TR of

estimated factors, and factor loadings. The results are reported in Figure 4. Figure 4 (a) shows that the GER estimator can obtain a high accuracy if the skewness is larger than 1.5 when α changes. For each skewness, the accuracy of the GER estimator decreases as α increases; this interactive effect becomes stronger when the skewness is milder. When the skewness is larger than one, the negative effect of α is small for the GER estimator. For the estimated factors and factor loadings, Figure 4 (b) and Figure 4 (c) show that the estimated factors and factor loadings only need a much smaller skewness strength to obtain a higher TR than the GER estimator needs. Approximately one skewness can support the high efficiency of the HFA factors and factor loadings. The factors and factor loadings also have different sensitivities for the skewness; this can be attributed to the asymptotic variance of HFA factor loadings affected by the higher-order cumulant of factors. As shown in Theorem 3, the variance of Λ becomes bigger as $\mathbf{C}_f^{(k)}$ decreases; thus, for each α, the TR of $\widehat{\Lambda}$ decreases as the skewness decreases. Conversely, the variance of F is not affected by $\mathbf{C}_f^{(k)}$; therefore, as observed in Figure 4 (b), when the skewness is larger than one, the TRs are close to one irrespective of the value of α.

$\sim$ Insert Figure 4 Here $\sim$

Second, we evaluate the sensitivity of the HFA estimation performance to the strength of the higher-order cumulant of u_{it}. We still follow the three-factor model in (37), changing the skewness of u_{it} from zero to two (for SGT distribution, we change η_j from 0 to 0.98) and α from zero to one in DGP1. The sample size is $(N, T) = (300, 500)$. As the HFA factors and the HFA factor loadings perform almost the same, we omit sensitivity analysis of the factor loadings here to conserve space. Both Figure 5 (a) and Figure 5 (b) show that the impact of u_{it}'s skewness is negligible for the HFA estimators and the GER estimator. This is expected because Proposition 1 implies that HFA still works for G_N with enough non-zero non-diagonal elements even when u_{it} is non-Gaussian. We give a special case that G_N is diagonal ($[G_N]_{jj} = j^{-0.544}$). At this time, the non-normality of u_{it} has a significant impact on the efficiency of HFA estimates. When the skewness of u_{it} increases, HFA loses efficiency in detecting and estimating the weak factors. This is expected since Proposition 1 implies that HFA cannot work well when N and T are comparably equal size if G_N is diagonal.

$\sim$ Insert Figure 5 Here $\sim$

Third, we evaluate the sensitivity of the HFA estimation accuracy to the decay rate of the spectrum ρ such that $\sigma_j(G_N) = j^{-\rho}$. We still follow the three-factor model in (37), changing ρ from zero to one and α from zero to one in DGP1. The sample size is $(N, T) = (300, 500)$. Figure 6 (a) and Figure 6 (b) show that the impact of ρ is important for the HFA estimators and the GER estimator. When $\rho \to 0$, HFA is inefficient to detect the weak factors. This is expected since both Theorem 1 and Theorem 2 implies that HFA cannot estimate the weak factors efficiently when N

and T are comparably equal size if $\text{tr}(G_N) = O(N)$. Increasing the dimension of T can effectively solve this problem, see Figure 6 (c) and (d).

$\sim$ Insert Figure 6 Here $\sim$

Overall, from our sensitivity analysis of the skewness of factors and errors, we can conclude that we need the moderate strength of the non-Gaussianity of factors to support the high accuracy of the GER estimator of the number of factors. For the estimation of the HFA factors and factor loadings, we only need mild non-Gaussianity of factors to support the high efficiency. When the skewness is larger than 1, this high efficiency exists in both strong and weak factor models ($\alpha \in [0, 1]$). Likewise, a moderate decay rate of the spectrum ρ ensures the HFA to work on the sample size where N and T are comparably equal. Additionally, the impact of the error's higher-order cumulants is negligible for the HFA estimators and the GER estimator if G_N has sufficient large number of non-zero non-diagonal elements.

7. Equity premium forecasting

In this section, we illustrate the usefulness of HFA for predicting the U.S. equity risk premium (ERP), defined by the excess return on the S&P 500 versus the U.S. Treasury Bill rate. It is computed as follows:

$$ERP_t = \log(1 + r_t^m) - \log(1 + r_t^f), \tag{41}$$

where r_t^m is the total return (including capital and dividend gains) on the S&P 500 portfolio — $r_t^m = (P_t - P_{t-1} + D_t)/P_{t-1}$ — where P_t is the S&P 500 index value, D_t are the dividends gained during the return period, and r_t^f denotes the U.S. Treasury Bill rate.

We evaluate the predictive value of the HFA factors extracted using FRED-MD. FRED-MD is a macroeconomic database of 134 monthly U.S. indicators. All series in FRED-MD are transformed to be stationary following the transformations described in McCracken & Ng (2016) and reject the null hypothesis of Augmented Dickey-Fuller test (Said & Dickey, 1984) at 5% significant level.[1] The series starts in January 1959 and ends in December 2018, with a total of 720 monthly observations.

The regression models used for forecasting take the form:

$$ERP_{t+1} = \mu + \beta(L)\widehat{f}_t + \gamma_h(L)ERP_t + \epsilon_{t+1}, \tag{42}$$

[1]Moreover, we delete the variables with more than 30 missing values (e.g., ACOGNO, TWEXAFEGSMTHx, UMCSENTx, and VXOCLSx) and use the MissForest algorithm of Stekhoven & Bühlmann (2012) to impute the remaining missing values. The dataset can be downloaded for free from the website `http://research.stlouisfed.org/econ/mccracken/sel/`. More details about the FRED-MD dataset can be found in McCracken & Ng (2016).

where ERP_{t+1} is the 1-step-ahead variable to be forecast, $\widehat{f}_t$ are the factors extracted using FRED-MD, and $\beta_h(L)$ and $\gamma_h(L)$ are finite order lag polynomials.

All models are estimated through a rolling-window estimation approach starting in January 1985 and comprising a 408 month out-of-sample. We consider two out-of-sample periods such that the first out-of-sample prediction begins in January 1985 and ends in October 2007 (Pre-crisis period). Further, the second out-of-sample period begins in November 2007, which is intended to evaluate the predictability during the crisis and recovery.

To have a better idea about the factor structure of the FRED-MD data set, we give the scree plots of the FRED-MD based on the covariance matrix (PCA), the third-order multi-cumulant (HFA3), and the fourth-order multi-cumulant (HFA4), respectively. Figure 7 intuitively shows the different factor structures presented by HFA and PCA. As shown in Figure 7, all scree plots show that one strong factor exists because we can observe a significant drawdown between the first and second singular values. It seems that no weak factors can be found from PCA's scree plot; conversely, two or three weak factors can be observed from the scree plots of HFA3 and HFA4 because another significant drawdown exists in Figure 7(b) and Figure 7(c).

$\sim$ Insert Figure 7 Here $\sim$

Assumption C requires normality of the error terms in the factor model. We validate this assumption on the HFA model-based residuals for the FRED-MD dataset using the normality test proposed by Bai & Ng (2005). Using a significance level of 5%, we find that after extracting four factors by HFA3 and HFA4, we reject the null hypothesis of normality for only 2.4% and 1.61% of the time series, respectively.[1] In addition, we measure the decay rate of the spectrum of the error terms in FRED-MD dataset. We find that after extracting four HFA3 and HFA4 factors, the estimated decay rate of the spectrum are 0.664 and 0.924, respectively. These results indicate that the main assumptions ensuring the reliability of the HFA approach are satisfied.

We further compare the predictive content of the three-factor structure: the first, second, and third ones are based on the GER, ER, and JJR criteria, respectively. For each factor structure, we consider HFA3, HFA4, JMCA, and PCA to extract the factors and compare their predictability.

For each month, we re-estimate the number of factors using the proposed GER estimators[2], Ahn & Horenstein (2013)'s ER estimator, and Jondeau et al. (2018)'s JJR approach. We find that the ER estimator always detects one factor that has strong influential power, and this is the same

[1] We also test in the rolling sample and the results are robust. See more details in section 9 of the Supplementary Appendix.

[2] We set $\widehat{R}^* = \max(\widehat{R}^{(3)}_{GER}, \widehat{R}^{(4)}_{GER})$ as the estimation of the number of non-Gaussian factors to avoid $\widehat{R}^{(3)}_{GER}$ underestimating it.

as the result in Figure 7(a). The GER estimator identifies between 1 and 4 factors, containing one strong factor and several weak factors; this is expected in Figure 7(b) and 7(c). The JJR approach always selects 4 factors.[1] Based on these results, we re-estimate the factors each month. The forecasting accuracy results are shown in Table 1. We use the predictive mean square error (MSE) and out-of-sample R-squared coefficient (R^2_{OOS}) to evaluate the performance. The R^2_{OOS} is defined as follows:

$$R^2_{OOS} = 100 \times (1 - \sum_{t=1}^{T^*}(\widehat{ERP}_t - ERP_t)^2 / \sum_{t=1}^{T^*}(\widehat{ERP}_t^{HA} - ERP_t)^2), \tag{43}$$

where T^* is the number of out-of-sample forecasts, and $\widehat{ERP}_t^{HA}$ is the historical average forecast, namely $\widehat{ERP}_{t+1|t}^{HA} = \frac{1}{t}\sum_{s=1}^{t} ERP_s$. To avoid bias, we determine the regression structure after the factor number is selected. To that end, we used BIC to select the number of autoregressive lags ($1 \leq p \leq 6$) and lags of the factors ($0 \leq m \leq 3$) over the rolling-window sample. Further, to compare the predictive difference between PCA and HFA, we use the Diebold & Mariano (1995)?s t-type test statistic to determine whether the forecast differences are statistically significant.

$\sim$ Insert Table 1 Here $\sim$

The results that assess the predictive power differences are reported in Table 1. The MSE ratio with an asterisk denotes that the DM test is significant at 10%, which implies that the prediction of this method is better than PCA. As can be observed, regardless of whether in Panels A, B, or C, the regression models based on HFA factors achieve the best rank compared to PCA for the three periods under consideration. The predictive difference between PCA and HFA can also be observed from the DM tests in Panel A and Panel B. Additionally, comparing Panel A with Panels B and C, it is clear that the forecast models with the GER criterion factor structure outperform the ER and the JJR criterion, especially in the crisis and post-crisis periods.

We further give the confidence interval of the S&P 500 equity risk premium of the factor-augmented regression with the HFA factors introduced in Section 5, where the asymptotic variance estimate of the equity risk premium is given in section 2 of the Supplementary Appendix. The number of factors is determined by the GER estimator. Figure 8 shows 95% interval prediction of HFA3 and HFA4 factor-augmented regressions, which cover the actual values of the equity risk premium well.

$\sim$ Insert Figure 8 Here $\sim$

[1] As the JJR approach always chooses the maximum number of factors $R_{\max}$ as the estimates, we constrain $R_{\max} = 4$ to simplify the regression form.

Following Welch & Goyal (2008), we use a graphical approach to illustrate the dynamics of the performance relative to the HFA factors versus the PCA factors during the prediction period. The cumulative sum-squared error (CSSE) between the PCA prediction and alternative prediction is used as the net-difference indicator. The CSSE of the HFA3 factor-based prediction is defined as follows:

$$CSSE_t = \sum_{s=1}^{t} (\widehat{ERP}_s^{PCA} - ERP_s)^2 - (\widehat{ERP}_s^{HFA3} - ERP_s)^2, \tag{44}$$

and similarly for other prediction methods. A positive CSSE means that the alternative model is performing better than the classical PCA prediction and vice versa. Figure 9 is the time series of CSSE for the HFA3, HFA4, and JMCA under the GER criterion factor structure, as already indicated in Panel A of Table 1. We smooth the time series curve with a six-month bandwidth. The shaded areas indicate the three largest drawdowns of the S&P 500 during the out-of-sample period. We can conclude three points from Figure 9. First, a general upward drift exists in HFA prediction. Second, we note that, as expected, it is especially in periods of crisis that the HFA outperforms the covariance-based approaches. This is explained by the more pronounced non-normality in the data in these periods. Third, the drift of HFA prediction remains positive over the most recent several decades. Overall, we can conclude that the HFA approach is more useful than the considered alternatives (PCA, JMCA) for extracting macro-economic factors for equity risk premium prediction.

∼ Insert Figure 9 Here ∼

8. Conclusion

This study developed a new framework for factor analysis based on the eigenvalue decomposition of the product between the higher-order multi-cumulant and its transpose. The proposed Higher-order multi-cumulant Factor Analysis (HFA) framework consistently estimates the number of factors, factors themselves, and factor loadings for high dimensional panel non-Gaussian data with an underlying weak factor structure. The rate of convergence and asymptotic distribution of the estimated factors and factor loadings are derived. Subsequently, we give the asymptotic properties of the factor-augmented regression by using the HFA factors. Our simulation studies confirm that the HFA estimators have good finite sample properties. Finally, we illustrate the usefulness of the HFA factors for forecasting the S&P 500 equity premium.

Acknowledgements

Wanbo Lu's research is sponsored by the National Science Foundation of China (71771187, 72011530149) and the Fundamental Research Funds for the Central Universities (JBK190602) in China. Guanglin Huang's research is sponsored by the China Postdoctoral Science Foundation (2023M742878). Kris Boudt received funding from the Research Foundation Flanders (G0G8320N). We are grateful to the editor, three anonymous referees, Dries Cornilly, and Serena Ng for helpful comments on a previous draft. We also thank participants at the econometrics seminar of Vrije Universiteit Amsterdam, the EC2 2020 conference at HEC Paris, and the HDTS 2021 workshop at Maastricht University.

References

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

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*, 135–171.

Bai, J., & Li, K. (2016). Maximum likelihood estimation and inference for approximate factor models of high dimension. *Review of Economics and Statistics*, *98*, 298–309.

Bai, J., Li, K. et al. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, *40*, 436–465.

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

Bai, J., & Ng, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business and Economic Statistics*, *23*, 49–60.

Bai, J., & Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, *74*, 1133–1150.

Bai, J., & Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, *176*, 18–29.

Bailey, N., Kapetanios, G., & Pesaran, M. H. (2021). Measurement of factor strength: Theory and practice. *Journal of Applied Econometrics*, *36*, 587–613.

Bonhomme, S., & Robin, J. M. (2009). Consistent noisy independent component analysis. *Journal of Econometrics*, *149*, 12–25.

Boudt, K., Cornilly, D., & Verdonck, T. (2020). Nearest comoment estimation with unobserved factors. *Journal of Econometrics*, *217*, 381 – 397.

Braun, M. L. (2006). Accurate error bounds for the eigenvalues of the kernel matrix. *The Journal of Machine Learning Research*, *7*, 2303–2328.

Cardoso, J. F., & Souloumiac, A. (1993). Blind beamforming for non-gaussian signals. *IEE proceedings F (Radar and Signal Processing)*, *140*, 362–370.

Chang, J., Guo, B., & Yao, Q. (2018). Principal component analysis for second-order stationary vector time series. *The Annals of Statistics*, *46*, 2094–2124.

Chen, L., Dolado, J. J., & Gonzalo, J. (2021). Quantile factor models. *Econometrica*, *89*, 875–910.

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*, 318–328.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*, 253–263.

Fan, J., Guo, J., & Zheng, S. (2022). Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, *117*, 852–861.

Freyaldenhoven, S. (2022). Factor models with local factorsdetermining the number of relevant factors. *Journal of Econometrics*, *229*, 80–102.

Granger, C. W. (1976). Tendency towards normality of linear combinations of random variables. *Metrika*, *23*, 237–248.

Jondeau, E., Jurczenko, E., & Rockinger, M. (2018). Moment component analysis: An illustration with international stock markets. *Journal of Business and Economic Statistics*, *36*, 576–598.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*, 455–500.

Li, Z., Ton, J.-F., Oglic, D., & Sejdinovic, D. (2021). Towards a unified analysis of random fourier features. *The Journal of Machine Learning Research*, *22*, 4887–4937.

Lu, W., & Huang, G. (2022). Estimating the higher-order co-moment with non-gaussian components and its application in portfolio selection. *Statistics*, *56*, 537–564.

McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business and Economic Statistics*, *34*, 574–589.

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

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, *168*, 244–258.

Risk, B. B., Matteson, D. S., & Ruppert, D. (2019). Linear non-gaussian component analysis via maximum likelihood. *Journal of the American Statistical Association*, *114*, 332–343.

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*, 599–607.

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest: Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*, 112–118.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*, 1167–1179.

Su, L., Shi, Z., & Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, *84*, 2215–2264.

Theodossiou, P. (1998). Financial data and the skewed generalized t distribution. *Management Science*, *44*, 1650–1661.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*, 1455–1508.

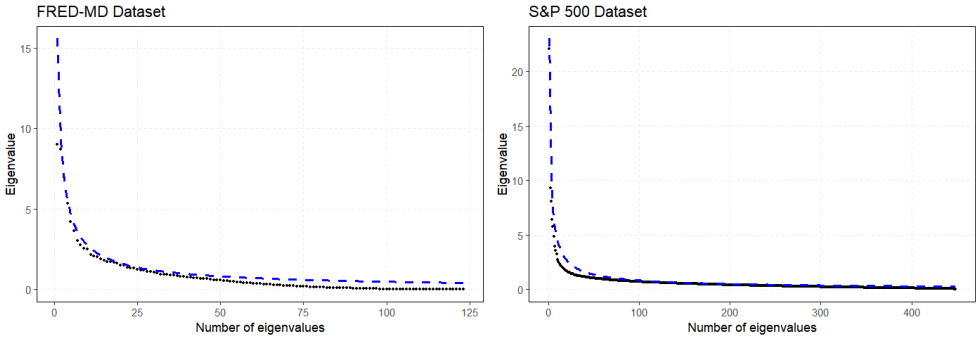
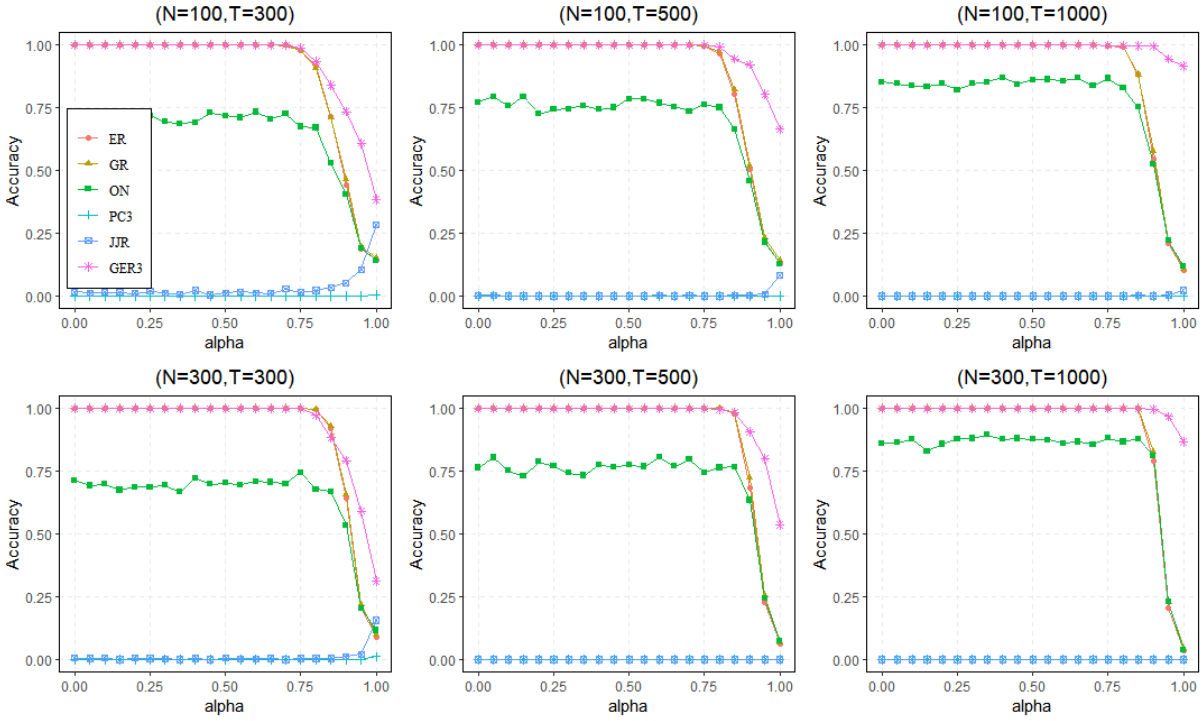
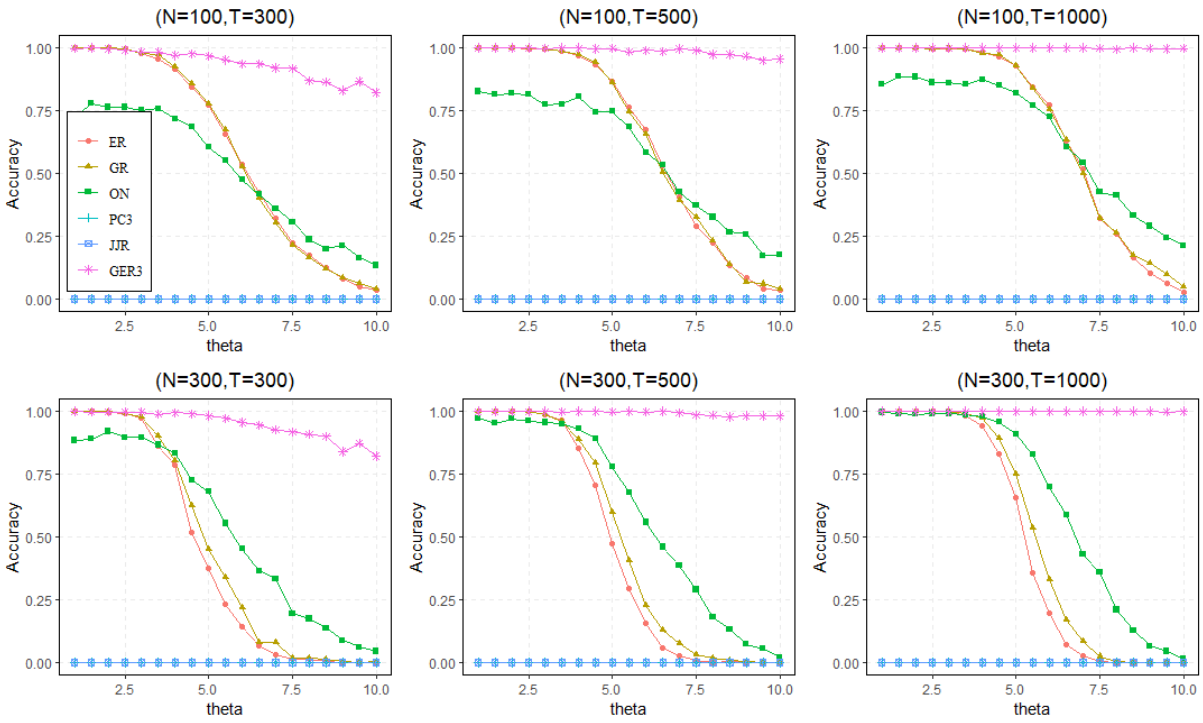


Figure 1: The eigenvalue spectrum of the error terms in two empirical datasets

Note: This figure reports the eigenvalue spectrum of the idiosyncratic errors after extracting the PCA factors. The black dots are the empirical eigenvalues of datasets. The blue dashed lines are the fitted distribution through a polynomial decaying function. The number of PCA factors is determined by Ahn & Horenstein (2013)'s ER estimator.



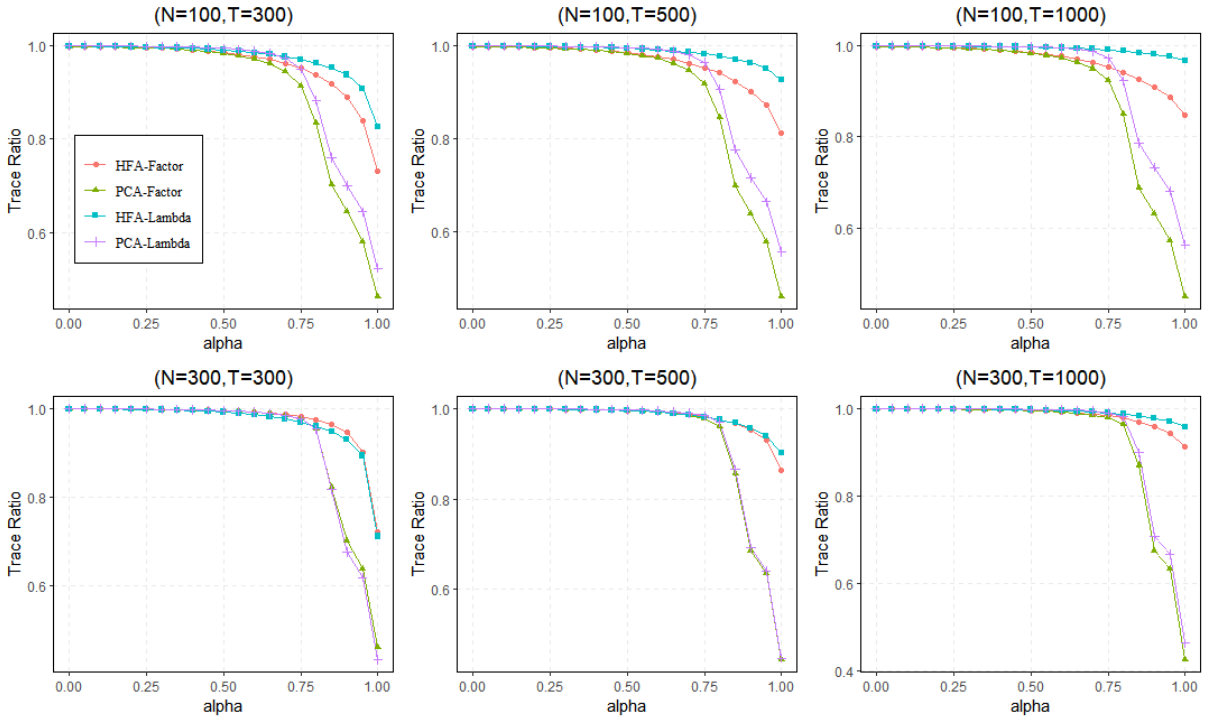
(a) DGP1



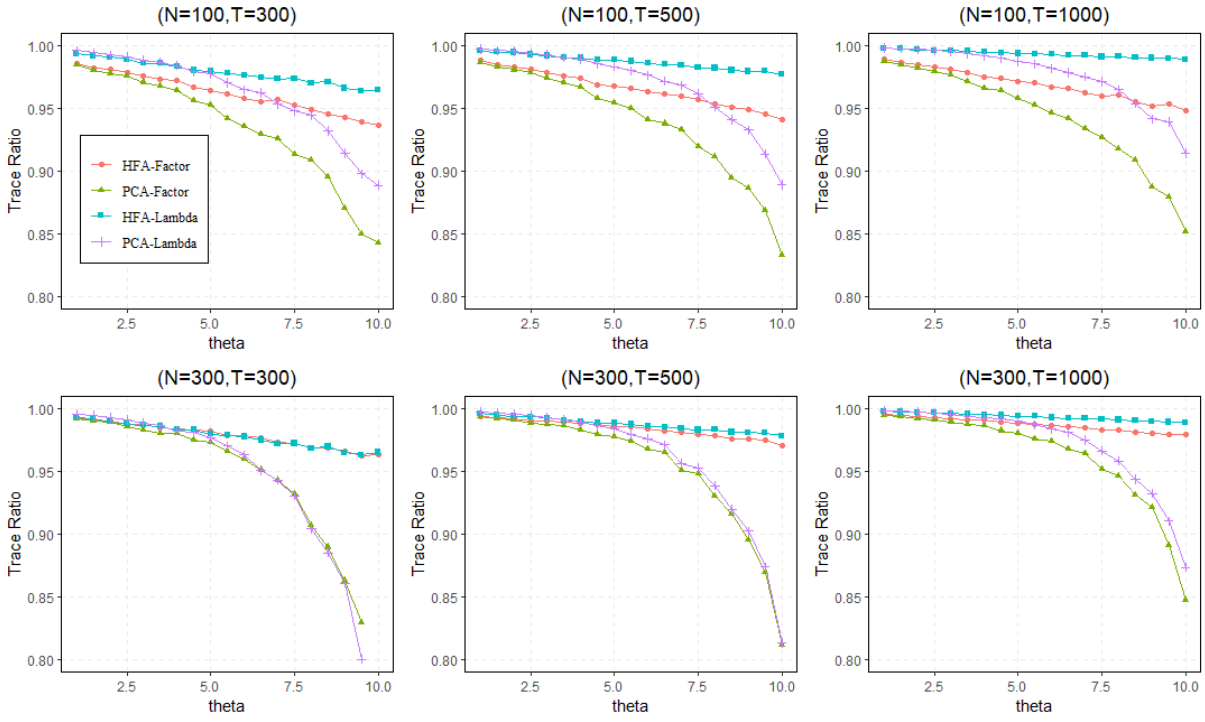
(b) DGP2

Figure 2: Accuracy of methods for selecting the number of factors

Note: This figure reports the proportion of the number of correctly selected non-Gaussian factors by six different methods: Bai & Ng (2002)'s PC3, Onatski (2010)'s ON estimator, Ahn & Horenstein (2013)'s ER and GR, Jondeau et al. (2018)'s JJR method, and the proposed GER estimator. The DGP1 and DGP2 follow a three-factor model in (37) and (38), respectively. α (in DGP1) and θ (in DGP2) control the factor strength. For each setting, we have 500 replications.



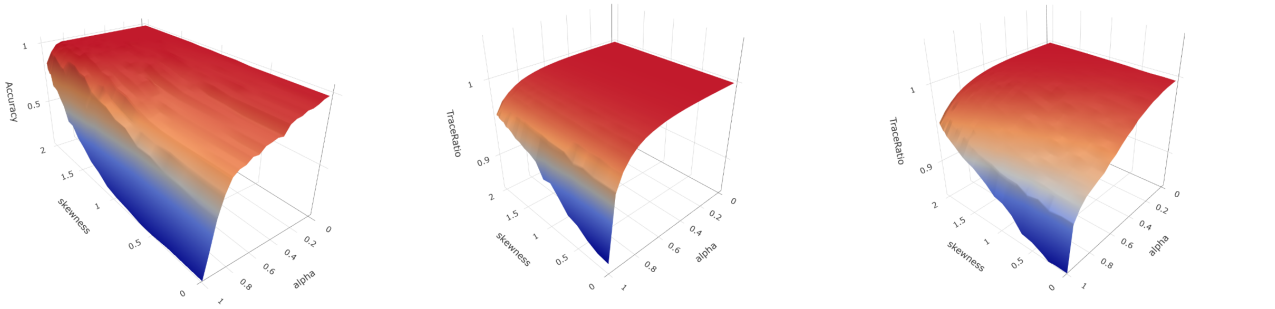
(a) DGP1



(b) DGP2

Figure 3: Accuracy of PCA and HFA estimators for the factors and loadings

Note: This figure reports the trace ratio (TR) of factors and factor loadings by the HFA estimators and PCA estimators. The DGP1 and DGP2 follow a three-factor model in (37) and (38), respectively. α (in DGP1) and θ (in DGP2) control the factor strength. For each estimation method, we report the median for 500 replications.



(a) Estimated factor number (b) Estimated factor (c) Estimated factor loading

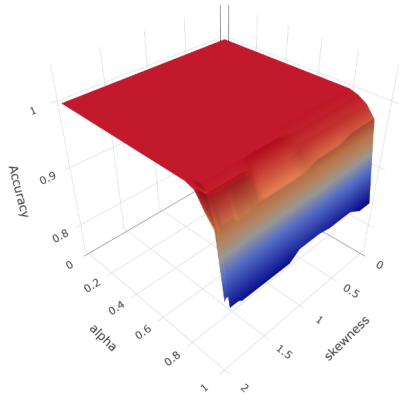
Figure 4: Impact of skewness of factors on the accuracy of the HFA estimator

Note: This figure reports the sensitivity analysis of the HFA estimators to the skewness of factors in a three-factor model. The DGP follows the model in (37), and the sample size is $(N, T) = (300, 500)$. The value α controls the factor strength and increases from 0 (strong factor model) to 1 (extreme weak factor model). The skewness of all three factors increases from 0 to 2. Figure (a) shows the accuracy of the GER estimator. Figures (b) and (c) show the average Trace Ratio of estimated factors and the corresponding factor loadings, respectively.

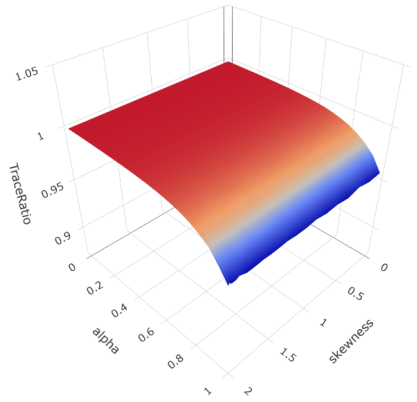
Table 1: Out-of-sample MSE for equity premium forecasting

	Pre-Crisis		Crisis & Post-Crisis		Full Sample	
	MSE	OS R^2	MSE	OS R^2	MSE	OS R^2
Panel A: GER criterion factor structure						
PCA	2.115	0.120	2.734	0.086	2.318	0.107
HFA3	2.074*	0.137	**2.676***	**0.105**	**2.272***	**0.125**
HFA4	**2.070***	**0.139**	2.719	0.091	2.283*	0.121
JMCA	2.110	0.122	2.734	0.086	2.315	0.109
Panel B: ER criterion factor structure						
PCA	2.113	0.121	2.730	0.087	2.316	0.108
HFA3	2.075*	0.137	2.786	0.069	2.308	0.111
HFA4	2.076*	0.137	2.787	0.068	2.309	0.111
JMCA	2.108	0.123	2.730	0.087	2.313	0.109
Panel C: JJR criterion factor structure						
PCA	2.177	0.094	2.847	0.048	2.397	0.077
HFA3	2.188	0.090	2.797	0.065	2.388	0.080
HFA4	2.136	0.112	2.874	0.039	2.378	0.084
JMCA	2.178	0.094	2.851	0.047	2.399	0.076

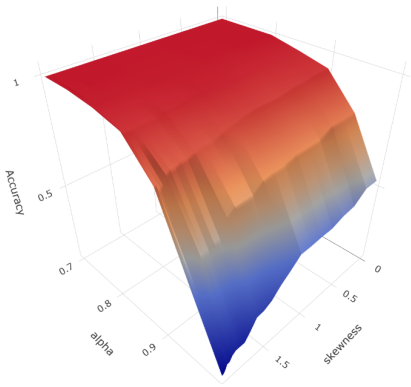
Note: This table reports the MSE and Out-of-Sample R^2 of four alternative 1-month-ahead forecasting methods for the equity premium (the best performing methods are shown in bold characters). The out-of-sample forecasting is implemented using rolling windows with 310 observations. The full forecasting evaluation period is from 1985-01 to 2018-12, the pre-crisis period is from 1985-01 to 2007-10, and the crisis plus post-crisis period is from 2007-11 to 2018-12. The MSE ratios with an asterisk denote that the left-sided DM test is significant at 10%, and the benchmark model of the DM test is PCA. Panels A, B, and C use the number of factors estimated by the GER, ER, and JJR estimator, respectively.



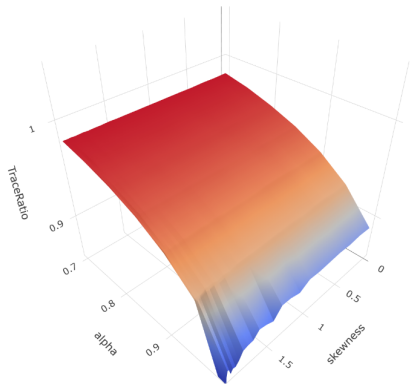
(a) Estimated factor number (diversified G_N)



(b) Estimated factor (diversified G_N)



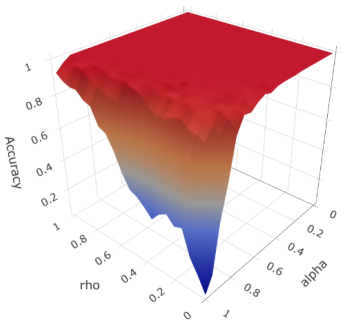
(c) Estimated factor number (diagonal G_N)



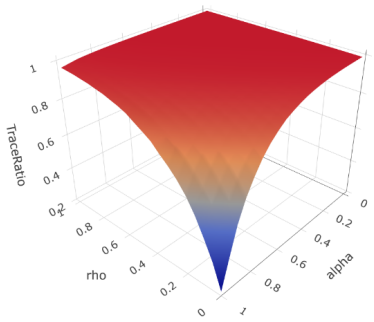
(d) Estimated factor (diagonal G_N)

Figure 5: Impact of skewness of u_{it} on the accuracy of the HFA estimator of the number of factors (left panel) and factor values (right panel)

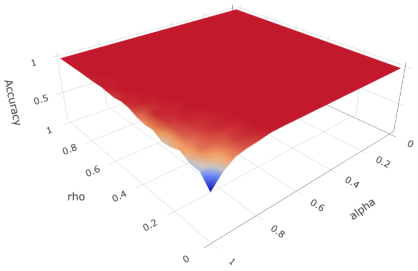
Note: This figure reports the sensitivity analysis of the HFA estimators concerning u_{it}'s skewness in a three-factor model with respect to two specific G_N calibrations: (i) sufficient large number of nonzero elements and (ii) diagonal matrix. The DGP follows the model in (37), and the sample size is $(N, T) = (300, 500)$. The value α controls the factor strength and increases from 0 (strong factor model) to 1 (extreme weak factor model). The skewness of u_{it} increases from 0 to 2. Figure (a) and (c) in left panel shows the accuracy of the GER estimator, while Figure (b) and (d) in right panel shows the average Trace Ratio of estimated HFA factors.



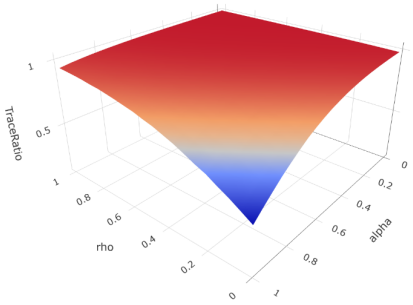
(a) Estimated factor number ($T = 500$)



(b) Estimated factor ($T = 500$)



(c) Estimated factor number ($T = 2000$)



(d) Estimated factor ($T = 2000$)

Figure 6: Impact of decay rate of the spectrum of G_N on the accuracy of the HFA estimator of the number of factors (left panel) and factor values (right panel)

Note: This figure reports the sensitivity analysis of the HFA estimators concerning $\text{tr}(G_N)$ with $\sigma_j(G_N) = j^{-\rho}$ in a three-factor model. The DGP follows the model in (37), and the sample size is $(N, T) = (300, 500), (300, 2000)$. The value α controls the factor strength and increases from 0 (strong factor model) to 1 (extreme weak factor model). Decay rate of the spectrum ρ increases from 0 to 1. Figure (a) and (c) in the left panel show the accuracy of the GER estimator, while Figure (b) and (d) in the right panel show the average Trace Ratio of estimated HFA factors.

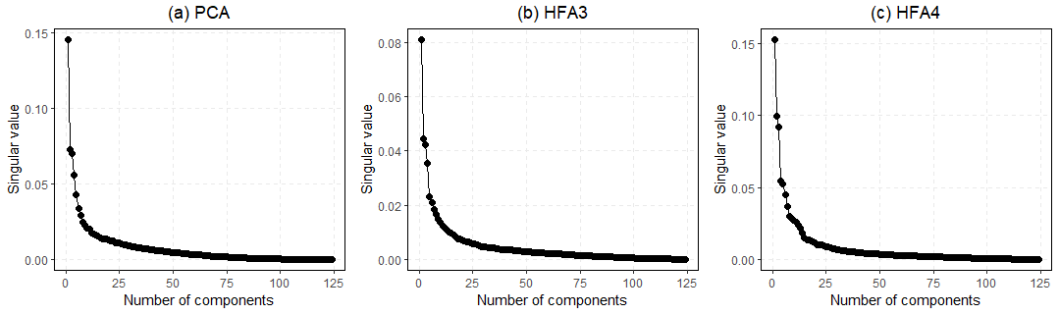
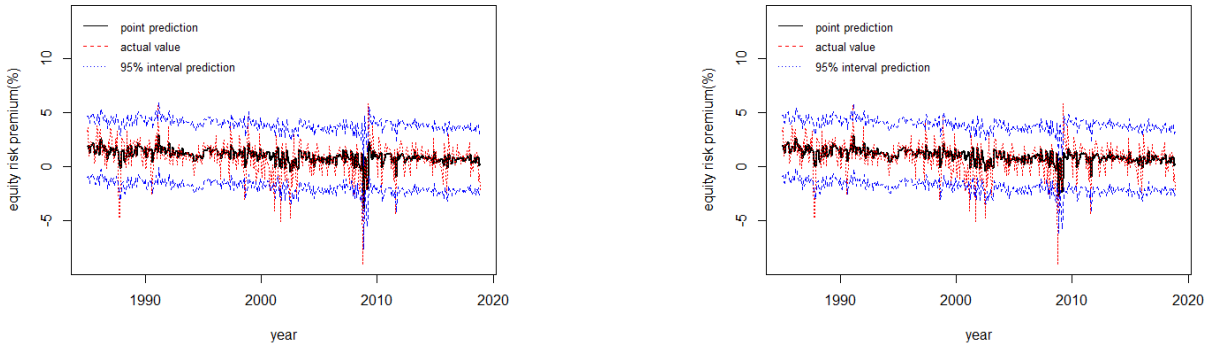


Figure 7: Scree plot of the FRED-MD data set

Note: This figure reports the singular values of the full sample FRED-MD database ($N = 124, T = 720$) based on PCA, HFA3, and HFA4, respectively. All series in FRED-MD are transformed to be stationary following the transformations described in McCracken & Ng (2016).



(a) HFA3 factor-augmented regression

(b) HFA4 factor-augmented regression

Figure 8: Interval prediction of the S&P 500 equity premium with HFA factors

Note: This figure reports the point prediction and the 95% interval prediction of the S&P 500 equity premium estimated by factor-augmented regression with the HFA factors. The S&P 500 equity premium is predicted on a monthly forecasting horizon from January 1985 until December 2018.

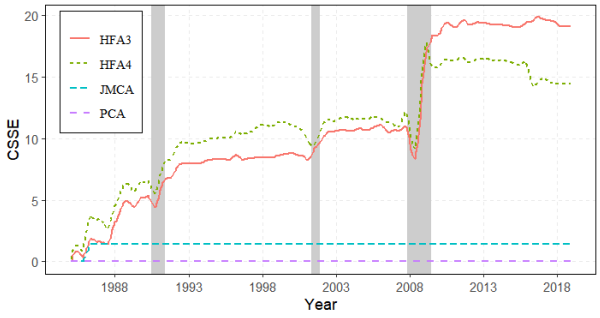


Figure 9: Out-of-sample CSSE for forecasting the S&P 500 equity premium

Note: This figure reports the cumulative sum-squared error of the HFA3, HFA4, and JMCA forecast based on the PCA forecast, which is defined in equation (44). The S&P 500 equity premium is predicted on a monthly forecasting horizon from January 1985 until December 2018. Shaded regions indicate the three largest drawdown periods of the S&P 500 during the out-of-sample period.

Supplementary appendix to: Estimation of non-Gaussian factors using higher-order multi-cumulants in weak factor models

Wanbo Lu[a], Guanglin Huang[b,d,*], Kris Boudt[c,d,e]

[a]*School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu 611130, China*
[b]*Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China*
[c]*Department of Economics, Universiteit Gent, 9000 Gent, Belgium*
[d]*Department of Business, Vrije Universiteit Brussel, 1050 Brussels, Belgium*
[e]*School of Business and Economics, Vrije Universiteit Amsterdam, 1081 Amsterdam, The Netherlands*

Abstract

In this supplementary appendix to the paper Lu et al. (2024) "Estimation of non-Gaussian factors using higher-order multi-cumulants in weak factor models", we first give the proofs of the lemmas, propositions and theorems. Second, we give the asymptotic results of the factor-augmented regressions based on HFA factors. Third, we provide a detailed elaboration of the loss function that is minimized by the alternating least squares algorithm in the presence of Gaussian factors. Fourth, we discuss how to handle non-stationarity of the data when implementing HFA. Fifth, we present an overview of several alternative factor estimation and selection approaches. Sixth, we present the computational aspects in HFA. Seventh, we provide more robustness checks regarding alternative specification of non-normality and additional application results. Eighth, we discuss the dampening of non-normality on Assumption C. Finally, a brief `R` tutorial for the proposed HFA method is provided. The proposed HFA estimators are available in the open source software `R` in the package `hofa`.

Keywords: Higher-order multi-cumulants, High-dimensional factor models, Weak factors, Consistency, R package
JEL: C100, C130, C510

*Corresponding author: SWUFE, Liutai Avenue 555, 611130 Chengdu, China. Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. (Email: huanggl@swufe.edu.cn)

1. Proofs

1.1. Lemmas

The following lemmas are useful to prove Theorem 1 - 3.

Lemma 1. *A and B are $p \times p$ positive definite and positive semi-definite matrices, respectively. Then, for any $j + k - 1 \leq i$, $\psi_i(AB) \leq \psi_j(A)\psi_k(B)$; $\psi_{p-j+1}(A)\psi_{p-k+1}(B) \leq \psi_{p-i+1}(AB)$.*

Proof: See Theorem 2.2 of Anderson & Gupta (1963).

Lemma 2. *Under Assumption A and B, denote $\mu_{NT,j}^{(k)} \equiv \sigma_j(N^{\frac{(\alpha-1)k}{2}}\Lambda\widetilde{\mathbf{C}}_f^{(k)}(\Lambda'^{\otimes(k-1)}))$ for $j = 1, 2, \ldots, R$. Then, for each $j = 1, 2, \ldots, R$, $p\lim_{(N,T)\to\infty}\mu_{NT,j}^{(k)} = \mu_j^{(k)}$, where $(\mu_j^{(k)})^2 = \psi_j(\Sigma_\Lambda\mathbf{C}_f^{(k)}\Sigma_\Lambda^{\otimes(k-1)}\mathbf{C}_f^{(k)\prime})$ and $0 < \mu_j^{(k)} < \infty$.*

Proof: Since $\mathbb{E}\|f_t\|^{2K} < \infty$ and $\|\lambda_i\| < \infty$ hold, it follows by Chebychev's Weak Law of Large Numbers that

$$p\lim_{(N,T)\to\infty}\psi_j(N^{(\alpha-1)k}\Lambda\widetilde{\mathbf{C}}_f^{(k)}(\Lambda'\Lambda)^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)\prime}\Lambda') = \psi_j(\Sigma_\Lambda\mathbf{C}_f^{(k)}\Sigma_\Lambda^{\otimes(k-1)}\mathbf{C}_f^{(k)\prime}),$$

where $\Sigma_\Lambda = \lim_{N\to\infty}\Lambda'\Lambda/N^{1-\alpha}$. By Assumption A and Lemma 1, we have

$$\begin{aligned}
\mu_{NT,j}^{(k)2} =& \psi_j(N^{(\alpha-1)k}\Lambda\widetilde{\mathbf{C}}_f^{(k)}(\Lambda'\Lambda)^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)\prime}\Lambda') \\
\leq& \psi_1(\Lambda'\Lambda/N^{1-\alpha})\psi_1((\Lambda'\Lambda/N^{1-\alpha})^{\otimes(k-1)})\psi_j(\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)\prime}) \\
\leq& \psi_1^k(\Lambda'\Lambda/N^{1-\alpha})\psi_j(\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)\prime}).
\end{aligned}$$

Moreover, by Lemma 1 again, we have

$$\begin{aligned}
\mu_{NT,j}^{(k)2} \geq& \psi_R(\Lambda'\Lambda/N^{1-\alpha})\psi_R((\Lambda'\Lambda/N^{1-\alpha})^{\otimes(k-1)})\psi_j(\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)\prime}) \\
\geq& \psi_R^k(\Lambda'\Lambda/N^{1-\alpha})\psi_j(\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)\prime}).
\end{aligned}$$

By Assumption A(ii) and Assumption B(ii), we have

$$\begin{aligned}
p\lim_{(N,T)\to\infty}\mu_{NT,j}^{(k)2} =& \mu_j^{(k)2} \leq p\lim_{(N,T)\to\infty}\psi_1^k(\Lambda'\Lambda/N^{1-\alpha})\psi_j(\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)\prime}) \leq (\nu_1)^k(\phi_j^{(k)})^2 < \infty, \\
p\lim_{(N,T)\to\infty}\mu_{NT,j}^{(k)2} =& \mu_j^{(k)2} \geq p\lim_{(N,T)\to\infty}\psi_R^k(\Lambda'\Lambda/N^{1-\alpha})\psi_j(\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)\prime}) \geq (\nu_R)^k(\phi_j^{(k)})^2 > 0,
\end{aligned}$$

for $j = 1, 2, \ldots, R$, where $\nu_1, \ldots, \nu_R$ are the eigenvalues of Σ_Λ.

Lemma 3. *Under Assumption C, for* $3 \le k \le K$, *we have*

$$\sigma_j(N^{\frac{(\alpha-1)k}{2}}\widetilde{\mathbf{C}}_e^{(k)}) = \sigma_j(G_N^{1/2})\mathrm{tr}(G_N)^{\frac{k-1}{2}}O_p(\sqrt{N^{(\alpha-1)k}\log N/T})$$

for $j = 1, \ldots, N$.

Proof: Since $E = UG_N^{1/2}$, we have $\widetilde{\mathbf{C}}_e^{(k)} = G_N^{1/2}\widetilde{\mathbf{C}}_u^{(k)}(G_N^{1/2})^{\otimes(k-1)}$. Notice that $\sigma_j(\widetilde{\mathbf{C}}_e^{(k)}) = \sqrt{\psi_j(\widetilde{\mathbf{C}}_e^{(k)}\widetilde{\mathbf{C}}_e^{(k)'})}$. For $\psi_j(\widetilde{\mathbf{C}}_e^{(k)}\widetilde{\mathbf{C}}_e^{(k)'})$, we have $\psi_j(\widetilde{\mathbf{C}}_e^{(k)}\widetilde{\mathbf{C}}_e^{(k)'}) \le \psi_j(G_N)\psi_1(\widetilde{\mathbf{C}}_u^{(k)}(G_N)^{\otimes(k-1)}\widetilde{\mathbf{C}}_u^{(k)'}) \le \psi_j(G_N^*)\psi_1(\widetilde{\mathbf{C}}_u^{(k)}(G_N^*)^{\otimes(k-1)}\widetilde{\mathbf{C}}_u^{(k)'})$, where G_N^* is a diagonal matrix of the eigenvalues of G_N. It follows that

$$\widetilde{\mathbf{C}}_u^{(k)}(G_N^*)^{\otimes(k-1)}\widetilde{\mathbf{C}}_u^{(k)'} = \sum_{i_1=1}^{N}\cdots\sum_{i_{k-2}=1}^{N} g_{i_1}^*\ldots g_{i_{k-2}}^*\widetilde{\mathbf{B}}_{u,i_1i_2\ldots i_{k-2}}^{(k)}G_N^*\widetilde{\mathbf{B}}_{u,i_1i_2\ldots i_{k-2}}^{(k)},$$

where $\widetilde{\mathbf{B}}_{u,i_1i_2\ldots i_{k-2}}^{(k)} \in \mathbb{R}^{N\times N}$ is the $(i_1i_2\ldots i_{k-2})$-th block matrix of $\widetilde{\mathbf{C}}_u^{(k)}$. To be specific, when $k = 3$, $\widetilde{\mathbf{C}}_u^{(3)}(G_N^*)^{\otimes(2)}\widetilde{\mathbf{C}}_u^{(3)'} = \sum_{i=1}^N g_i^*\widetilde{\mathbf{B}}_{u,i}^{(3)}G_N^*\widetilde{\mathbf{B}}_{u,i}^{(3)}$ and $\widetilde{\mathbf{C}}_u^{(3)} = [\widetilde{\mathbf{B}}_{u,1}^{(3)}, \ldots, \widetilde{\mathbf{B}}_{u,N}^{(3)}]$. When $k = 4$, $\widetilde{\mathbf{C}}_u^{(4)}(G_N^*)^{\otimes(3)}\widetilde{\mathbf{C}}_u^{(4)'} = \sum_{i=1}^N\sum_{j=1}^N g_i^*g_j^*\widetilde{\mathbf{B}}_{u,ij}^{(4)}G_N^*\widetilde{\mathbf{B}}_{u,ij}^{(4)}$ and $\widetilde{\mathbf{C}}_u^{(4)} = [\widetilde{\mathbf{B}}_{u,11}^{(4)}, \ldots, \widetilde{\mathbf{B}}_{u,1N}^{(4)}, \ldots, \widetilde{\mathbf{B}}_{u,NN}^{(4)}]$.

When $k = 3$, we have

$$\begin{aligned}
\psi_1(\widetilde{\mathbf{C}}_u^{(k)}(G_N^*)^{\otimes(k-1)}\widetilde{\mathbf{C}}_u^{(k)'}) &\le \sum_{i=1}^N g_i^*\psi_1(\widetilde{\mathbf{B}}_{u,i}^{(k)}G_N^*\widetilde{\mathbf{B}}_{u,i}^{(k)}) \\
&= \sum_{i=1}^N g_i^* \sup_{\|\mathbf{v}\|=1}\|\mathbf{v}'\widetilde{\mathbf{B}}_{u,i}^{(k)}G_N^*\widetilde{\mathbf{B}}_{u,i}^{(k)}\mathbf{v}\| \\
&= \sum_{i=1}^N g_i^*\sum_{j=1}^N g_j^* \sup_{\|\mathbf{v}\|=1}\left([\widetilde{\mathbf{B}}_{u,i}^{(k)}\mathbf{v}]_j\right)^2 \\
&\le \max_i \sup_{\|\mathbf{v}\|=1}\|\widetilde{\mathbf{B}}_{u,i}^{(k)}\mathbf{v}\|_\infty^2\sum_{i=1}^N\sum_{j=1}^N g_i^*g_j^*.
\end{aligned}$$

Notice that $\sum_{i=1}^N\sum_{j=1}^N g_i^*g_j^* = \mathrm{tr}(G_N)^2$. We thus only need to bound $\max_i \sup_{\|\mathbf{v}\|=1}\|\widetilde{\mathbf{B}}_{u,i}^{(k)}\mathbf{v}\|_\infty^2$. For $i = 1, \ldots, N$, let $[\widetilde{\mathbf{B}}_{u,i}^{(k)}]_{jk} = \xi_{ijk} = \frac{1}{T}\sum_{t=1}^T \xi_{ijk,t}$, where $\xi_{ijk,t} = u_{it}u_{jt}u_{kt}$. Since u_{it}, u_{jt} and u_{kt} are mutually independent and normal distributed, we have $\mathbb{E}(\xi_{ijk,t}) = 0$ and $\mathrm{Var}(\xi_{ijk,t}) \le \mathbb{E}(u_{it}^6) < \infty$. Therefore, by Theorem 3.3 and Lemma 2.4 of Saulis & Statulevicius (1991), for any $\|\mathbf{v}\| = 1$, we have $P(|\sum_{k=1}^N v_k\xi_{ijk,t}| > x) \le \exp(-c^*x^{\frac{2}{3}})$ with some universal positive constant c^*. Notice that the strong mixing coefficient satisfies $\bar{\alpha}_i(n) \le C_{\bar{\alpha}}\tau^n = C_{\bar{\alpha}}\exp(-\log(1/\tau)n)$. Now by Theorem 4.17

and Lemma 2.4 of Saulis & Statulevicius (1991), we have

$$\begin{aligned} P(\|\widetilde{\mathbf{B}}_{u,i}^{(k)}\mathbf{v}\|_\infty \geq \varepsilon) &= P(\max_{j\in[N]} \big|\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N} v_k \xi_{ijk,t}\big| \geq \varepsilon) \\ &\leq N \max_{j\in[N]} P(\big|\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N} v_k \xi_{ijk,t}\big| \geq \varepsilon) \\ &\leq N \max_{j\in[N]} \Big\{ \exp\big(-c^* T\varepsilon^2\big) + \exp\big(-c^*(T\varepsilon)^{\frac{2}{7}}\big)\Big\} \\ &\leq N \exp\big(-c^* T\varepsilon^2\big) \end{aligned}$$

for sufficient small ε. This implies $\max_i \sup_{\|\mathbf{v}\|=1} \|\widetilde{\mathbf{B}}_{u,i}^{(k)}\mathbf{v}\|_\infty \leq \sqrt{c^* \log(N/\delta)/T}$ with probability at least $1-\delta$. The bound of $\psi_1(\widetilde{\mathbf{C}}_u^{(k)}(G_N^*)^{\otimes(k-1)}\widetilde{\mathbf{C}}_u^{(k)'})$ for $k>3$ can be derived similarly. Consequently, we obtain

$$\sigma_j(N^{\frac{(\alpha-1)k}{2}}\widetilde{\mathbf{C}}_e^{(k)}) = \sigma_j(G_N^{1/2})\mathrm{tr}(G_N)^{\frac{k-1}{2}} O_p(\sqrt{N^{(\alpha-1)k}\log N/T}). \tag{1}$$

Lemma 4. *Under Assumptions A and B, for $3 \leq k \leq K$, it holds that*

$$\widetilde{\mathbf{C}}_x^{(k)} = \Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)} + \widetilde{\mathbf{C}}_e^{(k)} + \widetilde{\Pi}^{(k)},$$

where $\widetilde{\mathbf{C}}_x^{(k)}$, $\widetilde{\mathbf{C}}_f^{(k)}$ and $\widetilde{\mathbf{C}}_e^{(k)}$ are, respectively, the sample k-th order multi-cumulant of x_t, f_t and e_t, and $\widetilde{\Pi}^{(k)}$ contains the cross terms between Λf_t and e_t. To be specific, $\widetilde{\Pi}^{(k)} = \sum_{m=1}^{k-1}\widetilde{\Pi}_m^{(k)}$ with

$$\widetilde{\Pi}_m^{(k)} = \sum_{\binom{k}{m}\text{ permutation}} \frac{1}{T}\sum_{t=1}^{T} \mathbf{cum}\big(\underbrace{e_t, e_t, \ldots, \Lambda f_t}_{m \text{ of } \Lambda f_t\ ,\ k-m \text{ of } e_t}\big), \quad m = 1, \ldots, k-1,$$

where $\mathbf{cum}(e_t, e_t, \ldots, \Lambda f_t) \in \mathbb{R}^{N\times N^{k-1}}$ is the k-th order multi-cumulant of $\underbrace{\{e_t, e_t, \ldots, \Lambda f_t\}}_{m \text{ of } \Lambda f_t\ ,\ k-m \text{ of } e_t}$, and $\sum_{\binom{k}{m}\text{ permutation}}$ indicates the summation of $\binom{k}{m}$ permutation of $\{e_t, e_t, \ldots, \Lambda f_t\}$.[1] *Moreover, under Assumptions A – C, for $3 \leq k \leq K$, it holds that*

$$\sigma_1(\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\widetilde{\Pi}^{(k)}) \asymp \sigma_1(\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\widetilde{\Pi}_{k-1}^{(k)}).$$

[1] For example, when $k=3$ and $m=1$, we can write

$$\widetilde{\Pi}_1^{(3)} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}(e_t, e_t, \Lambda f_t) + \frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}(e_t, \Lambda f_t, e_t) + \frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}(\Lambda f_t, e_t, e_t).$$

Proof: Under Assumptions A and B, for $3 \leq k \leq K$, we have

$$\begin{aligned}
\widetilde{\mathbf{C}}_x^{(k)} &= \frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}\big(x_t,\ldots,x_t\big) \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}\big(\Lambda f_t+e_t,\ldots,\Lambda f_t+e_t\big) \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}\big(\Lambda f_t,\ldots,\Lambda f_t\big)+\frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}\big(e_t,\ldots,e_t\big)+\sum_{m=1}^{k-1}\sum_{\binom{k}{m}}\frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}\big(\underbrace{e_t,e_t,\ldots,\Lambda f_t}_{m \text{ of } \Lambda f_t\ ,\ k-m \text{ of } e_t}\big) \\
&= \Lambda\Big\{\frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}\big(f_t,\ldots,f_t\big)\Big\}(\Lambda'^{\otimes(k-1)})+\widetilde{\mathbf{C}}_e^{(k)}+\widetilde{\Pi}^{(k)} \\
&= \Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}+\widetilde{\mathbf{C}}_e^{(k)}+\widetilde{\Pi}^{(k)}.
\end{aligned}$$

Hence, the first part of Lemma 4 holds. Recall $e_t = G_N^{1/2}u_t$ under Assumption C. We have

$$\widetilde{\Pi}_m^{(k)} = \sum_{\binom{k}{m}\text{ permutation}}\frac{1}{T}\sum_{t=1}^{T}G_N^{1/2}\mathbf{cum}\big(\underbrace{u_t,u_t,\ldots,f_t}_{m \text{ of } f_t\ ,\ k-m \text{ of } u_t}\big)\big(\underbrace{G_N^{1/2}\otimes\ldots\otimes\Lambda'}_{m \text{ of } \Lambda'\ ,\ k-m-1 \text{ of } G_N^{1/2}}\big), \quad m=1,\ldots,k-1,$$

Following the proof of Lemma 3, let

$$\widetilde{\mathbf{C}}_{m,u,f}^{(k)} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{cum}\big(\underbrace{u_t,u_t,\ldots,f_t}_{m \text{ of } f_t\ ,\ k-m \text{ of } u_t}\big).$$

We have

$$\begin{aligned}
\sigma_1(\widetilde{\Pi}_m^{(k)}) &\asymp \sum_{\binom{k}{m}\text{ permutation}}\sigma_1\big(G_N^{1/2}\widetilde{\mathbf{C}}_{m,u,f}^{(k)}(G_N^{1/2}\otimes\ldots\otimes\Lambda')\big) \\
&\asymp \binom{k}{m}\sigma_1\big(G_N^{1/2}\widetilde{\mathbf{C}}_{m,u,f}^{(k)}(G_N^{1/2}\otimes\ldots\otimes\Lambda')\big).
\end{aligned}$$

Notice that $k \leq K$ is a finite number, it follows that $\sigma_1(\widetilde{\Pi}_m^{(k)}) \asymp \sigma_1\big(G_N^{1/2}\widetilde{\mathbf{C}}_{m,u,f}^{(k)}(G_N^{1/2}\otimes\ldots\otimes\Lambda')\big)$. Hence,

$$\begin{aligned}
\sigma_1^2(\widetilde{\Pi}_m^{(k)}) = \psi_1(\widetilde{\Pi}_m^{(k)}\widetilde{\Pi}_m^{(k)\prime}) &\asymp N^{(1-\alpha)m}\psi_1(G_N)\psi_1\big(\widetilde{\mathbf{C}}_{m,u,f}^{(k)}\big\{(G_N)^{\otimes(k-m-1)}\otimes(\Lambda'\Lambda/N^{1-\alpha})^{\otimes m}\big\}\widetilde{\mathbf{C}}_{m,u,f}^{(k)\prime}\big) \\
&\asymp N^{(1-\alpha)m}\psi_1(G_N)\psi_1\big(\widetilde{\mathbf{C}}_{m,u,f}^{(k)}\big\{(G_N^*)^{\otimes(k-m-1)}\otimes(\Sigma_\Lambda^*)^{\otimes m}\big\}\widetilde{\mathbf{C}}_{m,u,f}^{(k)\prime}\big),
\end{aligned}$$

where G_N^* and Σ_Λ^* are, respectively, the eigenvalue matrices of $G_N^{1/2}$ and $\Lambda'\Lambda/N^{1-\alpha}$. When $m = k-1$, the term $(G_N^*)^{\otimes(k-m-1)}$ is vanished, the rank of $\widetilde{\mathbf{C}}_{k-1,u,f}^{(k)}(\Sigma_\Lambda^*)^{\otimes(k-1)}\widetilde{\mathbf{C}}_{k-1,u,f}^{(k)'}$ is R^{k-1}, which is a finite number. Thus, Theorem 3.3 and Lemma 2.4 of Saulis & Statulevicius (1991) used in Lemma 3 cannot be applied, we can only derive $\psi_1\big(\widetilde{\mathbf{C}}_{k-1,u,f}^{(k)}(\Sigma_\Lambda^*)^{\otimes(k-1)}\widetilde{\mathbf{C}}_{k-1,u,f}^{(k)'}\big) \asymp O_p(N/T)$. When $m < k-1$, write $\widetilde{\mathbf{C}}_{m,u,f}^{(k)} = [\widetilde{\mathbf{B}}_{1\cdots1,m,u,f}^{(k)}, \ldots, \widetilde{\mathbf{B}}_{1\cdots N,m,u,f}^{(k)}, \ldots, \widetilde{\mathbf{B}}_{N\cdots N,m,u,f}^{(k)}]$, where $\widetilde{\mathbf{B}}_{i_1\cdots i_{k-2},m,u,f}^{(k)} \in \mathbb{R}^{N\times R}$ is the $i_1\cdots i_{k-2}$ block matrix of $\widetilde{\mathbf{C}}_{m,u,f}^{(k)}$. Notice that u_t and f_t are mutually independent by Assumption A(iii), we can follow Theorem 3.3 and Lemma 2.4 of Saulis & Statulevicius (1991) used in Lemma 3 to derive

$$\psi_1(\widetilde{\Pi}_m^{(k)}\widetilde{\Pi}_m^{(k)'}) \asymp \psi_1(G_N)\cdot \mathrm{tr}(G_N)^{k-m-1}\cdot O(N^{(1-\alpha)m})\cdot O_p(\log N/T).$$

Overall, when $k \geq 3$, we have

$$\psi_1(\widetilde{\Pi}_m^{(k)}\widetilde{\Pi}_m^{(k)'}) \asymp \begin{cases} \psi_1(G_N)\cdot N^{(1-\alpha)k}\cdot O_p(N^\alpha/T), & m = k-1; \\ \psi_1(G_N)\cdot \mathrm{tr}(G_N)^{k-m-1}\cdot N^{(1-\alpha)m}\cdot O_p(\log N/T), & 1 \leq m < k-1. \end{cases}$$

For the terms $1 \leq m < k-1$, (i) the dominated term is $\psi_1(\widetilde{\Pi}_1^{(k)}\widetilde{\Pi}_1^{(k)'})$ if $\mathrm{tr}(G_N) >> O(N^{1-\alpha})$, (ii) the dominated term is $\psi_1(\widetilde{\Pi}_{k-2}^{(k)}\widetilde{\Pi}_{k-2}^{(k)'})$ if $\mathrm{tr}(G_N) << O(N^{1-\alpha})$, (iii) all terms are $O_p(N^{(1-\alpha)(k-1)}\log N/T)$ if $\mathrm{tr}(G_N) \asymp O(N^{1-\alpha})$. For the case (ii) and (iii), it is obvious that $\psi_1(\widetilde{\Pi}_{k-1}^{(k)}\widetilde{\Pi}_{k-1}^{(k)'})$ dominates the remaining terms $\psi_1(\widetilde{\Pi}_m^{(k)}\widetilde{\Pi}_m^{(k)'})$ with $1 \leq m < k-1$. For the case (i), notice that by Lemma 3, $\psi_1(\widetilde{\mathbf{C}}_e^{(k)}\widetilde{\mathbf{C}}_e^{(k)'}) = \psi_1(G_N)\mathrm{tr}(G_N)^{k-1}O_p(N^\alpha/T) >> \psi_1(\widetilde{\Pi}_1^{(k)}\widetilde{\Pi}_1^{(k)'})$. Together with cases (i)-(iii), it follows that

$$\sigma_1(\mathbf{C}_e^{(k)}) + \sigma_1(\widetilde{\Pi}^{(k)}) \asymp \sigma_1(\mathbf{C}_e^{(k)}) + \sigma_1(\widetilde{\Pi}_{k-1}^{(k)}).$$

Lemma 5. *Under Assumption A – C, for $3 \leq k \leq K$, it holds that*

$$\sigma_{R+j}(N^{\frac{(\alpha-1)k}{2}}\widetilde{\mathbf{C}}_x^{(k)}) = \sigma_1(G_N^{1/2})\mathrm{tr}(G_N)^{\frac{k-1}{2}}O_p(\sqrt{N^{(\alpha-1)k}\log N/T}) + O_p(\sqrt{N^\alpha/T}),$$

for $j = 1, \ldots, N-R$ and $3 \leq k \leq K$.

Proof: By Lemma 4, we have

$$\widetilde{\mathbf{C}}_x^{(k)} = \Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)} + \widetilde{\mathbf{C}}_e^{(k)} + \widetilde{\Pi}^{(k)}, \quad k \geq 3.$$

Since $\mathrm{rank}(\Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}) = R$, $\tilde{\mu}_{NT,R+j}^{(k)} \leq \sigma_{R+j}(\Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}) + \sigma_1(\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\widetilde{\Pi}^{(k)}) = \sigma_1(\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\widetilde{\Pi}^{(k)})$. Lemma 4 implies that $\sigma_1(\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\widetilde{\Pi}^{(k)}) \asymp \sigma_1(\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\widetilde{\Pi}_{k-1}^{(k)})$. Notice that $\sigma_1(\widetilde{\Pi}_{k-1}^{(k)}) = N^{(1-\alpha)k/2}O_p(\sqrt{N^\alpha/T})$, then Lemma 3 and 4 imply the result.

Lemma 6. *Under Assumption A – C, for $j = 1, 2, \ldots, R$ and $3 \leq k \leq K$, if*

$$\mathrm{tr}(G_N) = o\big(N^{\frac{k}{k-1}(1-\alpha)} T^{\frac{1}{k-1}} (\log N)^{-\frac{1}{k-1}}\big)$$

and $N^\alpha/T = o(1)$, then we have

$$\sigma_j(N^{\frac{(\alpha-1)k}{2}} \widetilde{\mathbf{C}}_x^{(k)}) = \mu_{NT,j}^{(k)} + o_p(1),$$

where $\mu_{NT,j}^{(k)}$ is defined in Lemma 2.

Proof: Notice that $\tilde{\mu}_{NT,j}^{(k)} \leq \sigma_j(\Lambda \widetilde{\mathbf{C}}_f^{(k)} \Lambda'^{\otimes(k-1)}) + \sigma_1(\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\widetilde{\Pi}^{(k)}) = \mu_{NT,j}^{(k)} + \sigma_1(\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\widetilde{\Pi}^{(k)})$ for $j = 1, \ldots, R$. By Lemma 3, 4 and 5, when $\mathrm{tr}(G_N)^{k-1} N^{(\alpha-1)k} \log N/T = o(1)$ and $N^\alpha/T = o(1)$, the result follows.

Lemma 7. *Under Assumptions A – C, if $\mathrm{tr}(G_N) = o\big(N^{\frac{k}{k-1}(1-\alpha)} T^{\frac{1}{k-1}} (\log N)^{-\frac{1}{k-1}}\big)$ and $N^\alpha/T = o(1)$, we have*

$$p \lim_{(T,N)\to\infty} \frac{\Lambda' \widehat{\Lambda}^{(k)}}{N^{1-\frac{\alpha}{2}}} = Q^{(k)}. \tag{2}$$

The matrix $Q^{(k)}$ is invertible and is given by $Q^{(k)} = (\Psi^{(k)})^{-1/2} \Gamma^{(k)} (D^{(k)})^{1/2}$, where $\Psi^{(k)} = \mathbf{C}_f^{(k)} \Sigma_\Lambda^{\otimes(k-1)} \mathbf{C}_f^{(k)'}$, $D^{(k)} = diag(v_1, v_2, \ldots, v_R)$ are the eigenvalue of $(\Psi^{(k)})^{1/2} \Sigma_\Lambda (\Psi^{(k)})^{1/2}$, $\Sigma_\Lambda = \lim_{N\to\infty} \Lambda'\Lambda/N^{1-\alpha}$, and $\Gamma^{(k)}$ is corresponding eigenvector matrix such that $\Gamma^{(k)\prime} \Gamma^{(k)} = \mathbf{I}_R$.

Proof: Multiply the identity $N^{-(1-\alpha)k} \widetilde{\mathbf{C}}_x^{(k)} \widetilde{\mathbf{C}}_x^{(k)'} \widehat{\Lambda}^{(k)} = \widehat{\Lambda}^{(k)} \widetilde{D}_{NT}^{(k)}$ on both sides by $(\widetilde{\mathbf{C}}_f^{(k)} (\Lambda'\Lambda/N^{1-\alpha})^{\otimes(k-1)} \widetilde{\mathbf{C}}_f^{(k)'})^{1/2} \Lambda'/N^{1-\alpha/2}$ to obtain

$$(\widetilde{\mathbf{C}}_f^{(k)} (\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)} \widetilde{\mathbf{C}}_f^{(k)'})^{1/2} \frac{\Lambda'}{N^{1-\alpha/2}} (\frac{\widetilde{\mathbf{C}}_x^{(k)} \widetilde{\mathbf{C}}_x^{(k)'}}{N^{(1-\alpha)k}}) \widehat{\Lambda}^{(k)} = (\widetilde{\mathbf{C}}_f^{(k)} (\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)} \widetilde{\mathbf{C}}_f^{(k)'})^{1/2} \frac{\Lambda' \widehat{\Lambda}^{(k)}}{N^{1-\alpha/2}} \widetilde{D}_{NT}^{(k)}. \tag{3}$$

Notice that $\psi_j(\frac{\widetilde{\mathbf{C}}_x^{(k)} \widetilde{\mathbf{C}}_x^{(k)'}}{N^{(1-\alpha)k}}) = \psi_j(\Lambda \widetilde{\mathbf{C}}_f^{(k)} (\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)} \widetilde{\mathbf{C}}_f^{(k)'} \Lambda'/N^{1-\alpha}) + o_p(1)$ for $j = 1, 2, \ldots, N - R$ by Lemma 6. We can rewrite (3) as

$$\begin{aligned} &(\widetilde{\mathbf{C}}_f^{(k)} (\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)} \widetilde{\mathbf{C}}_f^{(k)'})^{1/2} (\frac{\Lambda'\Lambda}{N^{1-\alpha}}) (\widetilde{\mathbf{C}}_f^{(k)} (\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)} \widetilde{\mathbf{C}}_f^{(k)'}) (\frac{\Lambda' \widehat{\Lambda}^{(k)}}{N^{1-\alpha/2}}) \\ &= (\widetilde{\mathbf{C}}_f^{(k)} (\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)} \widetilde{\mathbf{C}}_f^{(k)'})^{1/2} (\frac{\Lambda' \widehat{\Lambda}^{(k)}}{N^{1-\alpha/2}}) \{\widetilde{D}_{NT}^{(k)} + o_p(1)\}. \end{aligned} \tag{4}$$

Let $\Psi_{NT}^{(k)} = \widetilde{\mathbf{C}}_f^{(k)} (\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)} \widetilde{\mathbf{C}}_f^{(k)'}$, then $B_{NT}^{(k)} = (\Psi_{NT}^{(k)})^{1/2} (\frac{\Lambda'\Lambda}{N^{1-\alpha}}) (\Psi_{NT}^{(k)})^{1/2}$, and $J_{NT}^{(k)} = (\Psi_{NT}^{(k)})^{1/2} (\frac{\Lambda'\widehat{\Lambda}}{N^{1-\alpha/2}})$.

Then we can rewrite (4) as

$$\{B_{NT}^{(k)} + o_p(1)\}J_{NT}^{(k)} = J_{NT}^{(k)}\widetilde{D}_{NT}^{(k)}. \tag{5}$$

Thus each column of $J_{NT}^{(k)}$ is the eigenvector of $B_{NT}^{(k)} + o_p(1)$ without standardization. Let $\widetilde{D}_{NT}^{(k)\dagger}$ be the eigenvalue matrix of $J_{NT}^{(k)\prime}J_{NT}^{(k)}$. Denote $\Gamma_{NT}^{(k)} = J_{NT}^{(k)}\widetilde{D}_{NT}^{(k)\dagger - \frac{1}{2}}$ so that each column of $\Gamma_{NT}^{(k)}$ has a unit length, and we have

$$\{B_{NT}^{(k)} + o_p(1)\}\Gamma_{NT}^{(k)} = \Gamma_{NT}^{(k)}\widetilde{D}_{NT}^{(k)\dagger}. \tag{6}$$

Thus $\Gamma_{NT}^{(k)}$ is the eigenvector matrix of $B_{NT}^{(k)} + o_p(1)$. Note that $B_{NT}^{(k)} + o_p(1)$ converges to $B^{(k)} = (\Psi^{(k)})^{1/2}\Sigma_\Lambda(\Psi^{(k)})^{1/2}$ by Assumption A and B. In addition, the eigenvalues of $B_{NT}^{(k)}$ is distinct by Lemma 2. Following Bai (2003)'s proofs of Proposition 1, there exists a unique eigenvector matrix $\Gamma^{(k)}$ of $B^{(k)}$ such that $\|\Gamma^{(k)} - \Gamma_{NT}^{(k)}\| = o_p(1)$. From

$$\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha/2}} = (\Psi_{NT}^{(k)})^{-1/2}\Gamma_{NT}^{(k)}(\widetilde{D}_{NT}^{(k)\dagger})^{1/2}, \tag{7}$$

we have

$$\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha/2}} \to (\Psi^{(k)})^{-1/2}\Gamma^{(k)}(D^{(k)})^{1/2} \tag{8}$$

by Assumptions A - B and $\widetilde{D}_{NT}^{(k)\dagger} \to_p D^{(k)}$.

Lemma 8. *Define the rotation matrix*

$$H^{(k)} = \widetilde{\mathbf{C}}_f^{(k)}(\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)\prime}(\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha}})(\widetilde{D}_{NT}^{(k)})^{-1}. \tag{9}$$

Under Assumptions A – C and if $\mathrm{tr}(G_N) = o\big(N^{\frac{k}{k-1}(1-\alpha)}T^{\frac{1}{k-1}}(\log N)^{-\frac{1}{k-1}}\big)$ *and* $N^\alpha/T = o(1)$,

$$(\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N})'H^{(k)} = \mathbf{I}_R + o_p(1), \tag{10}$$

where $\widetilde{D}_{NT}^{(k)}$ *is* $R \times R$ *diagonal matrix of the first* R *largest eigenvalues of* $\frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime}$ *in decreasing order.*

Proof: We have

$$
\begin{aligned}
(\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N})'H^{(k)} =&(\frac{\widehat{\Lambda}^{(k)\prime}\Lambda}{N})\widetilde{\mathbf{C}}_f^{(k)}(\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}(\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha}})(\widetilde{D}_{NT}^{(k)})^{-1}\\
=&(\frac{\widehat{\Lambda}^{(k)\prime}}{\sqrt{N}})(\frac{1}{N^{1-\alpha}}\Lambda\widetilde{\mathbf{C}}_f^{(k)}(\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}\Lambda')(\frac{\widehat{\Lambda}^{(k)}}{\sqrt{N}})\widetilde{D}_{NT}^{(k)-1}.
\end{aligned}
$$

Notice that $\psi_j(\frac{\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'}}{N^{(1-\alpha)k}}) = \psi_j(\frac{1}{N^{1-\alpha}}\Lambda\widetilde{\mathbf{C}}_f^{(k)}(\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}\Lambda') + o_p(1)$ for $j = 1, 2, \ldots, N-R$ by Lemma 6. It follows that

$$
(\frac{\widehat{\Lambda}^{(k)\prime}}{\sqrt{N}})(\frac{1}{N^{1-\alpha}}\Lambda\widetilde{\mathbf{C}}_f^{(k)}(\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}\Lambda')(\frac{\widehat{\Lambda}^{(k)}}{\sqrt{N}}) = \widetilde{D}_{NT}^{(k)} + o_p(1).
$$

Therefore,

$$
\begin{aligned}
(\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha}})'H^{(k)} &= \{\widetilde{D}_{NT}^{(k)} + o_p(1)\}\widetilde{D}_{NT}^{(k)-1}\\
&= \mathbf{I}_R + o_p(1).
\end{aligned}
$$

1.2. Proof of Theorem 1

The proof of Theorem 1 follows the same strategy as that in Ahn & Horenstein (2013). The main results are based on the asymptotic properties of $\widetilde{\mu}_{NT,r}^{(k)}$: If $\lim_{T\to\infty}\widetilde{\mathbf{C}}_f^{(k)}$ is full rank R, we have $\widetilde{\mu}_{NT,r}^{(k)}/\widetilde{\mu}_{NT,r+1}^{(k)} = O_p(1)$ for $r \neq R$ and $\widetilde{\mu}_{NT,R}^{(k)}/\widetilde{\mu}_{NT,R+1}^{(k)} \to \infty$ as $(N,T) \to \infty$. Indeed, while the ratio of the R-th and $(R+1)$-th singular values of $\widetilde{\mathbf{C}}_x^{(k)}$ diverge to infinity, all other ratios of the two adjacent singular values are asymptotically bounded. All lemmas used to prove the theorem can be found in this Supplementary Appendix.

By Lemma 6, for $j = 1, 2, \ldots, R-1$, if $\mathrm{tr}(G_N) = o\big(N^{\frac{k}{k-1}(1-\alpha)}T^{\frac{1}{k-1}}(\log N)^{-\frac{1}{k-1}}\big)$ and $N^\alpha/T = o(1)$ holds, then we have $\tilde{\mu}_{NT,j}^{(k)}/\tilde{\mu}_{NT,j+1}^{(k)} = \mu_{NT,j}^{(k)}/\mu_{NT,j+1}^{(k)} + o_p(1) = O_p(1)$. Subsequently, by Lemma 5, $\tilde{\mu}_{NT,R}^{(k)}/\tilde{\mu}_{NT,R+1}^{(k)} = \big(\mu_{NT,R}^{(k)} + o_p(1)\big)/o_p(1) \to \infty$ as $(N,T) \to \infty$. By Lemma 3, for $j = 1, \ldots, N-R-1$, $\tilde{\mu}_{NT,R+j}^{(k)}/\tilde{\mu}_{NT,R+j+1}^{(k)} = O_p(1)$. Thus, we have

$$
p\lim_{(N,T)\to\infty}\tilde{\mu}_{NT,j}^{(k)}/\tilde{\mu}_{NT,j+1}^{(k)} = \begin{cases} \infty & , \quad j = R, \\ O(1) & , \quad j \neq R. \end{cases} \tag{11}
$$

Subsequently, the consistency of the GER estimator follows from (11).

1.3. Proof of Theorem 2

Our proof follows the same strategy as that in Bai (2003) for the estimation of Λ using $\widetilde{\mathbf{C}}_x^{(2)} = \frac{1}{T}X'X$. The proof mainly comprises the following steps. First, we state the identity between the proposed estimator $\widehat{\Lambda}^{(k)}$ and the sample higher-order multi-cumulant $\widetilde{\mathbf{C}}_x^{(k)}$. Subsequently, we generalize the rotation matrix $H^{(k)}$ in Bai (2003) and give the explicit form of the estimation errors of factor loadings $\widehat{\Lambda}^{(k)} - \Lambda H^{(k)}$ and factors $\widehat{F}^{(k)} - F(H^{(k)'})^{-1}$. Finally, we give the convergence rate of the estimators based on that of the estimation errors.

Let $\widetilde{D}_{NT}^{(k)}$ be an $R \times R$ diagonal matrix of the first R largest eigenvalues of $\frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'}$ in decreasing order. By the definition of eigenvectors and eigenvalues, we have $\frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'}\widehat{\Lambda}^{(k)} = \widehat{\Lambda}^{(k)}\widetilde{D}_{NT}^{(k)}$ or

$$\widehat{\Lambda}^{(k)} = \frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'}\widehat{\Lambda}^{(k)}(\widetilde{D}_{NT}^{(k)})^{-1}. \tag{12}$$

Recall the rotation matrix in (9),

$$H^{(k)} = \widetilde{\mathbf{C}}_f^{(k)}(\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}(\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha}})(\widetilde{D}_{NT}^{(k)})^{-1}.$$

By Lemma 6, $\|\widetilde{D}_{NT}^{(k)}\| \leq \sum_{j=1}^{R}\tilde{\mu}_{NT,j}^{(k)2} = O_p(1)$. Under Assumptions A and B, we have $\|\frac{\Lambda'\Lambda}{N^{1-\alpha}}\| = O(1)$, $\|\widetilde{\mathbf{C}}_f^{(k)}\| = O_p(1)$ and $\|\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha/2}}\| = O_p(1)$. Subsequently, $\|H^{(k)}\| \leq N^{\frac{\alpha}{2}}\|\widetilde{\mathbf{C}}_f^{(k)}\|^2 \cdot \|\frac{\Lambda'\Lambda}{N^{1-\alpha}}\|^{k-1} \cdot \|\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha/2}}\| \cdot \|\widetilde{D}_{NT}^{(k)-1}\| = O_p(N^{\frac{\alpha}{2}})$.

Therefore, the estimation error between $\widehat{\Lambda}^{(k)}$ and Λ is

$$\begin{aligned}\widehat{\Lambda}^{(k)} - \Lambda H^{(k)} &= \frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'}\widehat{\Lambda}^{(k)}(\widetilde{D}_{NT}^{(k)})^{-1} - \Lambda\widetilde{\mathbf{C}}_f^{(k)}(\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}(\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha}})(\widetilde{D}_{NT}^{(k)})^{-1} \\ &= \sqrt{N}\Big\{\frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'} - \frac{1}{N^{(1-\alpha)k}}\Lambda\widetilde{\mathbf{C}}_f^{(k)}(\Lambda'\Lambda)^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}\Lambda'\Big\}(\frac{\widehat{\Lambda}^{(k)}}{\sqrt{N}})(\widetilde{D}_{NT}^{(k)})^{-1}.\end{aligned}$$

Notice that $AA' - BB' = (A-B)A' + B(A-B)'$ for two matrices $A, B \in \mathbb{R}^{p\times q}$, we have

$$\big\|\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)'} - \Lambda\widetilde{\mathbf{C}}_f^{(k)}(\Lambda'\Lambda)^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}\Lambda'\big\| \leq \sigma_1\big(\widetilde{\mathbf{C}}_x^{(k)} - \Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}\big)\big(\big\|\widetilde{\mathbf{C}}_x^{(k)}\big\| + \big\|\Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}\big\|\big).$$

Since $\big\|\frac{1}{N^{(1-\alpha)k/2}}\widetilde{\mathbf{C}}_x^{(k)}\big\| \asymp \big\|\frac{1}{N^{(1-\alpha)k/2}}\Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}\big\| = O_p(1)$. It suffice to bound $\frac{1}{N^{(1-\alpha)k/2}}\sigma_1(\widetilde{\mathbf{C}}_x^{(k)} - \Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)})$. Notice that

$$\frac{1}{N^{(1-\alpha)k/2}}\widetilde{\mathbf{C}}_x^{(k)} - \frac{1}{N^{(1-\alpha)k/2}}\Lambda\widetilde{\mathbf{C}}_f^{(k)}(\Lambda')^{\otimes(k-1)} = \frac{1}{N^{(1-\alpha)k/2}}\widetilde{\mathbf{C}}_e^{(k)} + \frac{1}{N^{(1-\alpha)k/2}}\widetilde{\Pi}^{(k)},$$

where $\widetilde{\Pi}^{(k)}$ is defined in Lemma 4. By Lemma 3 and 4, we have

$$\begin{aligned}\frac{1}{N^{(1-\alpha)k/2}}\sigma_1(\widetilde{\mathbf{C}}_x^{(k)} - \Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}) &\leq \sigma_1(\frac{1}{N^{(1-\alpha)k/2}}\widetilde{\mathbf{C}}_e^{(k)}) + \sigma_1(\frac{1}{TN^{(1-\alpha)k/2}}\widetilde{\Pi}^{(k)}) \\ &= O_p(\sqrt{\mathrm{tr}(G_N)^{k-1}N^{(\alpha-1)k}\log N/T}) + O_p(\sqrt{N^\alpha/T}).\end{aligned}$$

Together with $\|\frac{\widehat{\Lambda}^{(k)}}{\sqrt{N}}\| = 1$ and $\|(\widetilde{D}_{NT}^{(k)})^{-1}\| = O_p(1)$, the average convergence rate of $\widehat{\Lambda}^{(k)}$ is

$$\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(k)} - \Lambda H^{(k)}\| = O_p(\sqrt{\mathrm{tr}(G_N)^{k-1}N^{(\alpha-1)k}\log N/T}) + O_p(\sqrt{N^\alpha/T}). \tag{13}$$

For the estimated factors $\widehat{F}^{(k)}$, we have $\widehat{F}^{(k)} = X\widehat{\Lambda}^{(k)}/N = F\Lambda'\widehat{\Lambda}^{(k)}/N + E\Lambda H^{(k)}/N + E(\widehat{\Lambda}^{(k)} - \Lambda H^{(k)})/N$. Lemma 8 implies that $\Lambda'\widehat{\Lambda}^{(k)}/N \to (H^{(k)'})^{-1}$. Therefore, when $(N,T) \to \infty$, substituting $(H^{(k)'})^{-1}$ for $\Lambda'\widehat{\Lambda}^{(k)}/N$ to obtain

$$\frac{1}{\sqrt{T}}\|\widehat{F}^{(k)} - F(H^{(k)'})^{-1}\| \leq \frac{1}{\sqrt{T}}\|E\Lambda H^{(k)}/N\| + \frac{1}{\sqrt{T}}\|E(\widehat{\Lambda}^{(k)} - \Lambda H^{(k)})/N\|.$$

The first part holds $\frac{1}{\sqrt{T}}\|E\Lambda H^{(k)}/N\| \leq \frac{1}{\sqrt{N^{1+\alpha}}}(\frac{1}{\sqrt{T}}\|E\Lambda/\sqrt{N^{1-\alpha}}\|)\cdot O_p(N^{\alpha/2}) = O_p(\frac{1}{\sqrt{N}})O_p(1)$. The second part holds as $\frac{1}{\sqrt{T}}\|E(\widehat{\Lambda}^{(k)} - \Lambda H^{(k)})/N\| \leq \sigma_1(\frac{1}{\sqrt{NT}}E)\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(k)} - \Lambda H^{(k)}\| = O_p(\frac{1}{\sqrt{TN^{1-\alpha}}}) + O_p(\sqrt{\frac{\mathrm{tr}(G_N)^{k-1}N^{(\alpha-1)k}\log N}{NT}})$. Overall, we have

$$\frac{1}{\sqrt{T}}\|\widehat{F}^{(k)} - F(H^{(k)\prime})^{-1}\| = O_p(\frac{1}{\sqrt{N}}) + O_p(\frac{1}{\sqrt{TN^{1-\alpha}}}) + O_p(\sqrt{\frac{\mathrm{tr}(G_N)^{k-1}N^{(\alpha-1)k}\log N}{NT}}).$$

1.4. Proof of Theorem 3

The proof of Theorem 3 follows the same strategy as that of Theorem 2. For each $\widehat{\lambda}_i^{(k)}$ and $\widehat{f}_t^{(k)}$, we give the explicit form of the estimation error. Subsequently, we find the dominant term and use the sandwich formula to derive the asymptotic distribution of the proposed estimators. We use the same notation as that in the proof of Theorem 2. We denote $[A]_i$ as the i-th column of matrix A.

To derive the distribution of $\widehat{\lambda}_i^{(k)}$, we need define a new rotation matrix as follows

$$\bar{H}^{(k)} = \frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}\widetilde{\mathbf{C}}_x^{(k)\prime}\widehat{\Lambda}^{(k)}(\widetilde{D}_{NT}^{(k)})^{-1}. \tag{14}$$

Notice that

$$\begin{aligned}\bar{H}^{(k)} &= \frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}\big(\Lambda'^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)'}\Lambda' + \widetilde{\mathbf{C}}_e^{(k)'} + \widetilde{\Pi}^{(k)}\big)\widehat{\Lambda}^{(k)}(\widetilde{D}_{NT}^{(k)})^{-1} \\ &= H^{(k)} + \frac{1}{N^{(1-\alpha)k}}\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}\big(\widetilde{\mathbf{C}}_e^{(k)'} + \widetilde{\Pi}^{(k)}\big)\widehat{\Lambda}^{(k)}(\widetilde{D}_{NT}^{(k)})^{-1} \\ &= H^{(k)} + o_p(1),\end{aligned}$$

provided that $\mathrm{tr}(G_N) = o\big(N^{\frac{k}{k-1}(1-\alpha)}T^{\frac{1}{k-1}}(\log N)^{-\frac{1}{k-1}}\big)$ and $N^\alpha/T = o(1)$. First, we derive the distribution of $\widehat{\lambda}_i^{(k)}$. By the definitions of $\bar{H}^{(k)}$ and equation (12), we have

$$\widehat{\lambda}_i^{(k)} - \bar{H}^{(k)\prime}\lambda_i = \widetilde{D}_{NT}^{(k)-1}\Big(\frac{1}{\sqrt{N}}\widehat{\Lambda}^{(k)'}\Big)\Big(\frac{1}{N^{(1-\alpha)k/2}}\widetilde{\mathbf{C}}_x^{(k)}\Big)\frac{\sqrt{N}}{N^{(1-\alpha)k/2}}[\widetilde{\mathbf{C}}_x^{(k)'} - (\Lambda)^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)}\Lambda']_i. \tag{15}$$

Notice that $[\widetilde{\mathbf{C}}_x^{(k)'} - \Lambda^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)}\Lambda']_i \asymp [\widetilde{\Pi}_{k-1}^{(k)'} + \widetilde{\mathbf{C}}_e^{(k)'}]_i$ by Lemma 4. By Lemma 3, we have $\sigma_1(\frac{1}{N^{(1-\alpha)k/2}}\widetilde{\mathbf{C}}_e^{(k)'}) = \mathrm{tr}(G_N)^{\frac{k-1}{2}}O_p(\sqrt{N^{(\alpha-1)k}\log N/T})$. By Lemma 4, we have $\sigma_1(\frac{1}{N^{(1-\alpha)k/2}}\widetilde{\Pi}_{k-1}^{(k)'}) \asymp \sigma_1(\frac{1}{N^{(1-\alpha)k/2}}\Lambda^{\otimes(k-1)}\frac{1}{T}\bar{\mathcal{H}}_f^{(k)\prime}E) = O_p(\sqrt{N^\alpha/T})$. If $\mathrm{tr}(G_N) = o(N^{\frac{k}{k-1}-\alpha}(\log N)^{-\frac{1}{k-1}})$, then $\frac{1}{N^{(1-\alpha)k/2}}\widetilde{\mathbf{C}}_e^{(k)'}$ is negligible after multiplying $\sqrt{TN^{-\alpha}}$. Now (15) can be rewritten as

$$\begin{aligned}\widehat{\lambda}_i^{(k)} - \bar{H}^{(k)\prime}\lambda_i &= \widetilde{D}_{NT}^{(k)-1}\Big(\frac{1}{\sqrt{N}}\widehat{\Lambda}^{(k)'}\Big)\Big(\frac{\widetilde{\mathbf{C}}_x^{(k)}}{N^{(1-\alpha)k/2}}\Big)\frac{\sqrt{N}}{N^{(1-\alpha)k/2}}\Lambda^{\otimes(k-1)}\frac{1}{T}\bar{\mathcal{H}}_f^{(k)\prime}e_i + o_p\Big(\sqrt{\frac{N^\alpha}{T}}\Big) \\ &= \widetilde{D}_{NT}^{(k)-1}\Big(\frac{1}{\sqrt{N}}\widehat{\Lambda}^{(k)'}\Big)\Big(\frac{\Lambda\widetilde{\mathbf{C}}_f^{(k)}\Lambda'^{\otimes(k-1)}}{N^{(1-\alpha)k/2}}\Big)\frac{\sqrt{N}}{N^{(1-\alpha)k/2}}\Lambda^{\otimes(k-1)}\frac{1}{T}\bar{\mathcal{H}}_f^{(k)\prime}e_i + o_p\Big(\sqrt{\frac{N^\alpha}{T}}\Big) \\ &= (\widetilde{D}_{NT}^{(k)})^{-1}\Big(\frac{\widehat{\Lambda}^{(k)'}\Lambda}{N^{1-\alpha}}\Big)\widetilde{\mathbf{C}}_f^{(k)}\Big(\frac{\Lambda'\Lambda}{N^{1-\alpha}}\Big)^{\otimes(k-1)}\frac{1}{T}\bar{\mathcal{H}}_f^{(k)\prime}e_i + o_p\Big(\sqrt{\frac{N^\alpha}{T}}\Big).\end{aligned} \tag{16}$$

Now multiply both sides of the equation (16) by $\sqrt{TN^{-\alpha}}$ to obtain

$$\sqrt{TN^{-\alpha}}(\widehat{\lambda}_i^{(k)} - \bar{H}^{(k)\prime}\lambda_i) = (\widetilde{D}_{NT}^{(k)})^{-1}\Big(\frac{\widehat{\Lambda}^{(k)'}\Lambda}{N^{1-\alpha/2}}\Big)\widetilde{\mathbf{C}}_f^{(k)}\Big(\frac{\Lambda'\Lambda}{N^{1-\alpha}}\Big)^{\otimes(k-1)}\frac{1}{\sqrt{T}}\bar{\mathcal{H}}_f^{(k)\prime}e_i + o_p(1). \tag{17}$$

When $(N,T) \to \infty$, $N^\alpha/T \to 0$ and $\mathrm{tr}(G_N) = o(N^{\frac{k}{k-1}-\alpha}(\log N)^{-\frac{1}{k-1}})$, we have $\widetilde{\mathbf{C}}_f^{(k)} \to \mathbf{C}_f^{(k)}$, $\widetilde{D}_{NT}^{(k)} \to D^{(k)}$, $\frac{\widehat{\Lambda}^{(k)'}\Lambda}{N^{1-\alpha/2}} \to Q^{(k)'}$ and $\frac{\Lambda'\Lambda}{N^{1-\alpha}} \to \Sigma_\Lambda$. By Assumption D(i), we know that $\frac{1}{\sqrt{T}}\bar{\mathcal{H}}_f^{(k)\prime}e_i \xrightarrow{d} \mathcal{N}(0,\Theta_i^{(k)})$. Overall, following Proposition 2 of Bai & Ng (2023), substituting $H^{(k)}$ for $\bar{H}^{(k)}$, we have

$$\sqrt{TN^{-\alpha}}(\widehat{\lambda}_i^{(k)} - H^{(k)\prime}\lambda_i) \xrightarrow{d} \mathcal{N}\big(0,(D^{(k)})^{-1}Q^{(k)'}\mathbf{C}_f^{(k)}(\Sigma_\Lambda)^{\otimes k-1}\Theta_i^{(k)}(\Sigma_\Lambda)^{\otimes k-1}\mathbf{C}_f^{(k)'}Q^{(k)}(D^{(k)})^{-1}\big). \tag{18}$$

To derive the limit distribution of $\widehat{f}_t^{(k)}$, we follow Bai (2003)'s proof of Theorem 2. From $\widehat{F}^{(k)} =$

$X\widehat{\Lambda}^{(k)}/N$ and $X = F\Lambda' + E$, we have $\widehat{f}_t^{(k)} = N^{-1}\widehat{\Lambda}^{(k)\prime}\Lambda f_t + N^{-1}(\widehat{\Lambda}^{(k)} - \Lambda H^{(k)})'e_t + N^{-1}H^{(k)\prime}\Lambda' e_t$. Using $\Lambda'\widehat{\Lambda}^{(k)}/N \to (H^{(k)\prime})^{-1}$ in Lemma 8, we obtain

$$\widehat{f}_t^{(k)} - (H^{(k)})^{-1}f_t = \frac{1}{N}H^{(k)\prime}\Lambda' e_t + \frac{1}{N}(\widehat{\Lambda}^{(k)} - \Lambda H^{(k)})'e_t. \tag{19}$$

The last term holds $\|\frac{1}{N}(\widehat{\Lambda}^{(k)} - \Lambda H^{(k)})'e_t\| \asymp \frac{1}{\sqrt{T}}\|\frac{1}{N}E(\widehat{\Lambda}^{(k)} - \Lambda H^{(k)})\| \leq \sigma_1(\frac{1}{\sqrt{NT}}E)\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(k)} - \Lambda H^{(k)}\| = O_p(\frac{1}{\sqrt{TN^{1-\alpha}}}) + O_p(\sqrt{\frac{\mathrm{tr}(G_N)^{k-1}N^{(\alpha-1)k}\log N}{NT}})$. Therefore, if we have $N^\alpha/T \to 0$ and $\mathrm{tr}(G_N) = o\big(T^{\frac{1}{k-1}}N^{\frac{k}{k-1}(1-\alpha)}(\log N)^{-\frac{1}{k-1}}\big)$, the last term $\frac{1}{N}(\widehat{\Lambda}^{(k)} - \Lambda H^{(k)})'e_t$ is negligible when we multiply by $\sqrt{N}$. Equation (19) implies

$$\sqrt{N}\big(\widehat{f}_t^{(k)} - (H^{(k)})^{-1}f_t\big) = (\frac{1}{\sqrt{N^\alpha}}H^{(k)\prime})(\frac{1}{\sqrt{N^{1-\alpha}}}\Lambda' e_t) + o_p(1). \tag{20}$$

Together with Assumption D and $\frac{1}{\sqrt{N^\alpha}}H^{(k)} \to (Q^{(k)})^{-1}$, we have

$$\sqrt{N}(\widehat{f}_t^{(k)} - (H^{(k)})^{-1}f_t) \xrightarrow{d} \mathcal{N}(0, (Q^{(k)\prime})^{-1}\Phi_t(Q^{(k)})^{-1}). \tag{21}$$

1.5. Proof of Proposition 1

Recall that $G_N = LG_N^*L'$, then $\mathbf{C}_e^{(k)} = (LG_N^{*1/2}L')\mathbf{C}_u^{(k)}(LG_N^{*1/2}L')^{\otimes(k-1)}$. Since $\sigma_1^2(\mathbf{C}_e^{(k)}) = \psi_1(\mathbf{C}_e^{(k)}\mathbf{C}_e^{(k)\prime})$, we only need to study $\psi_1(\mathbf{C}_e^{(k)}\mathbf{C}_e^{(k)\prime})$. Let $\bar{\mathbf{C}}_u^{(k)} = L'\mathbf{C}_u^{(k)}(L)^{\otimes(k-1)}$. Notice that

$$\begin{aligned}\psi_1(\mathbf{C}_e^{(k)}\mathbf{C}_e^{(k)\prime}) &= \psi_1(G_N^*L'\mathbf{C}_u^{(k)}(L)^{\otimes(k-1)}(G_N^*)^{\otimes(k-1)}(L')^{\otimes(k-1)}\mathbf{C}_u^{(k)\prime}L)\\ &= \psi_1(G_N^*\bar{\mathbf{C}}_u^{(k)}(G_N^*)^{\otimes(k-1)}\bar{\mathbf{C}}_u^{(k)\prime}).\end{aligned}$$

(i) When $k = 3$, write $\mathbf{C}_u^{(3)} = [\mathbf{B}_{u,1}^{(3)}, \mathbf{B}_{u,2}^{(3)}, \ldots, \mathbf{B}_{u,N}^{(3)}]$, where $\mathbf{B}_{u,i}^{(3)}$ has all elements equal to zero, except the i-th diagonal element, which equals $\kappa_{u,i}^{(3)}$. Analogously, write $\bar{\mathbf{C}}_u^{(3)} = [\bar{\mathbf{B}}_{u,1}^{(3)}, \bar{\mathbf{B}}_{u,2}^{(3)}, \ldots, \bar{\mathbf{B}}_{u,N}^{(3)}]$, we have

$$\begin{aligned}\psi_1(\mathbf{C}_e^{(3)}\mathbf{C}_e^{(3)\prime}) &\leq \psi_1(G_N^*)\psi_1(\bar{\mathbf{C}}_u^{(3)}(G_N^*)^{\otimes(2)}\bar{\mathbf{C}}_u^{(3)\prime})\\ &\leq \psi_1(G_N^*)\sum_{i=1}^N g_i^*\psi_1(\bar{\mathbf{B}}_{u,i}^{(3)}G_N^*\bar{\mathbf{B}}_{u,i}^{(3)}).\end{aligned}$$

Note that $\bar{\mathbf{B}}_{u,i}^{(3)} = L'(\sum_{j=1}^N l_{j,i}\mathbf{B}_{u,j}^{(3)})L$ with $\sum_{j=1}^N l_{j,i}\mathbf{B}_{u,j}^{(3)} = \mathrm{diag}(l_{1,i}\kappa_{u,1}^{(3)}, \ldots, l_{N,i}\kappa_{u,N}^{(3)})$. Therefore,

$$\begin{aligned}
\psi_1(\bar{\mathbf{B}}_{u,i}^{(3)} G_N^* \bar{\mathbf{B}}_{u,i}^{(3)}) &= \psi_1(L' \sum_{j=1}^N l_{j,i}\mathbf{B}_{u,j}^{(3)} L G_N^* L' \sum_{j=1}^N l_{j,i}\mathbf{B}_{u,j}^{(3)} L) \\
&\leq \psi_1(\sum_{j=1}^N l_{j,i}\mathbf{B}_{u,j}^{(3)} \sum_{j=1}^N l_{j,i}\mathbf{B}_{u,j}^{(3)}) \psi_1(G_N^*) \\
&\leq \psi_1(G_N^*) \max_{j\in[N]} |l_{j,i}|^2 \max_{j\in[N]} |\kappa_{u,j}^{(3)}|^2.
\end{aligned}$$

Recall the number of nonzero elements in $(l_{1,i}, \ldots, l_{j,i})^\top$ is $\mathcal{G}_i$ and $\sqrt{\mathcal{G}_i}|l_{j,i}| = O(1)$ if $l_{j,i} \neq 0$. It follows that $\max_{j\in[N]} |l_{j,i}|^2 \asymp \mathcal{G}_i^{-1}$. Overall, we have

$$\begin{aligned}
\psi_1(\mathbf{C}_e^{(3)} \mathbf{C}_e^{(3)'}) &\leq \psi_1^2(G_N^*)\mathrm{tr}(G_N) \max_{i\in[N]} \max_{i\in[N]} |l_{j,i}|^2 |\kappa_u^{(3)}|^2 \\
&\lesssim \psi_1^2(G_N^*)\mathrm{tr}(G_N)\mathcal{G}^{-1} |\kappa_u^{(3)}|^2,
\end{aligned}$$

where $\kappa_u^{(3)} = \max_{j\in[N]} |\kappa_{u,j}^{(3)}|$ and $\mathcal{G} = \min_i \mathcal{G}_i$.

(ii) When $k = 4$, write $\mathbf{C}_u^{(4)} = [\mathbf{B}_{u,1,1}^{(4)}, \ldots, \mathbf{B}_{u,1,N}^{(4)}, \ldots, \mathbf{B}_{u,N,1}^{(4)}, \ldots, \mathbf{B}_{u,N,N}^{(4)}]$, where $\mathbf{B}_{u,i,j}^{(4)}$ has all elements equal to zero, except the i-th diagonal element of $\mathbf{B}_{u,i,i}^{(4)}$, which equals $\kappa_{u,i}^{(4)}$. Analogously, write $\bar{\mathbf{C}}_u^{(4)} = [\bar{\mathbf{B}}_{u,1,1}^{(4)}, \ldots, \bar{\mathbf{B}}_{u,1,N}^{(4)}, \ldots, \bar{\mathbf{B}}_{u,N,1}^{(4)}, \ldots, \bar{\mathbf{B}}_{u,N,N}^{(4)}]$, we have

$$\begin{aligned}
\psi_1(\mathbf{C}_e^{(4)} \mathbf{C}_e^{(4)'}) &\leq \psi_1(G_N^*)\psi_1(\bar{\mathbf{C}}_u^{(4)} (G_N^*)^{\otimes(3)} \bar{\mathbf{C}}_u^{(4)'}) \\
&\leq \psi_1(G_N^*) \sum_{j=1}^N \sum_{i=1}^N g_i^* g_j^* \psi_1(\bar{\mathbf{B}}_{u,i,j}^{(4)} G_N^* \bar{\mathbf{B}}_{u,i,j}^{(4)}).
\end{aligned}$$

Note that $\bar{\mathbf{B}}_{u,i,j}^{(4)} = L'(\sum_{p=1}^N \sum_{q=1}^N l_{p,i} l_{q,j} \mathbf{B}_{u,p,q}^{(4)})L$ with

$$\sum_{p=1}^N \sum_{q=1}^N l_{p,i} l_{q,j} \mathbf{B}_{u,p,q}^{(4)} \;=\; \mathrm{diag}(l_{1,i}l_{1,j}\kappa_{u,1}^{(4)}, \ldots, l_{N,i}l_{N,j}\kappa_{u,N}^{(4)}).$$

Therefore,

$$\begin{aligned}\psi_1(\bar{\mathbf{B}}^{(4)}_{u,i,j}G^*_N\bar{\mathbf{B}}^{(4)}_{u,i,j}) &= \psi_1\big(L'\sum_{p=1}^N\sum_{q=1}^N l_{p,i}l_{q,j}\mathbf{B}^{(4)}_{u,p,q}LG^*_NL'\sum_{p=1}^N\sum_{q=1}^N l_{p,i}l_{q,j}\mathbf{B}^{(4)}_{u,p,q}L\big)\\ &\le \psi_1(\sum_{p=1}^N\sum_{q=1}^N l_{p,i}l_{q,j}\mathbf{B}^{(4)}_{u,p,q}\sum_{p=1}^N\sum_{q=1}^N l_{p,i}l_{q,j}\mathbf{B}^{(4)}_{u,p,q})\psi_1(G^*_N)\\ &\le \psi_1(G^*_N)\max_{p\in[N]}|l_{p,i}|^2|l_{p,j}|^2\max_{p\in[N]}|\kappa^{(4)}_{u,p}|^2.\end{aligned}$$

Overall, we have

$$\begin{aligned}\psi_1(\mathbf{C}^{(4)}_e\mathbf{C}^{(4)'}_e) &\le \psi_1^2(G^*_N)\mathrm{tr}(G_N)^2\max_{i\in[N]}\max_{p\in[N]}|l_{p,i}|^2|l_{p,j}|^2\max_{p\in[N]}|\kappa^{(4)}_{u,p}|^2\\ &\lesssim \psi_1^2(G^*_N)\mathrm{tr}(G_N)^2\mathcal{G}^{-2}|\kappa^{(4)}_u|^2,\end{aligned}$$

where $\kappa^{(4)}_u = \max_{p\in[N]}|\kappa^{(4)}_{u,p}|$ and $\mathcal{G} = \min_i \mathcal{G}_i$.

(iii) When $k > 4$, we have

$$\psi_1(\mathbf{C}^{(k)}_e\mathbf{C}^{(k)'}_e) \le \psi_1(G^*_N)\psi_1(\bar{\mathbf{C}}^{(k)}_u(G^*_N)^{\otimes(k-1)}\bar{\mathbf{C}}^{(k)'}_u).$$

The matrix $\bar{\mathbf{C}}^{(k)}_u(G^*_N)^{\otimes(k-1)}\bar{\mathbf{C}}^{(k)}_u = \sum_{i_1=1}^N\cdots\sum_{i_{k-2}=1}^N g^*_{i_1}\cdots g^*_{i_{k-2}}\bar{\mathbf{B}}^{(k)}_{u,i_1i_2\ldots i_{k-2}}G^*_N\bar{\mathbf{B}}^{(k)}_{u,i_1i_2\ldots i_{k-2}}$, where $\bar{\mathbf{B}}^{(k)}_{u,i_1i_2\ldots i_{k-2}} \in \mathbb{R}^{N\times N}$ is the $(i_1i_2\ldots i_{k-2})$-th block matrix of $\bar{\mathbf{C}}^{(k)}_u$. Notice that

$$\begin{aligned}\bar{\mathbf{B}}^{(k)}_{u,i_1\ldots i_{k-2}} &= L'\Big(\sum_{\ell_1=1}^N\cdots\sum_{\ell_{k-2}=1}^N l_{\ell_1,i_1}\cdots l_{\ell_{k-2},i_{k-2}}\mathbf{B}^{(k)}_{u,\ell_1\ldots\ell_{k-2}}\Big)L\\ &= L'\mathrm{diag}(l_{1,i_1}\cdots l_{1,i_{k-2}}\kappa^{(k)}_{u,1},\ldots,l_{N,i_1}\cdots l_{N,i_{k-2}}\kappa^{(k)}_{u,N})L.\end{aligned}$$

Analogously, we have

$$\psi_1(\bar{\mathbf{B}}^{(k)}_{u,i_1i_2\ldots i_{k-2}}G^*_N\bar{\mathbf{B}}^{(k)}_{u,i_1i_2\ldots i_{k-2}}) \le \psi_1(G^*_N)\max_{j\in[N]}(|l_{j,i_1}|^2\cdots|l_{j,i_{k-2}}|^2)\max_{j\in[N]}|\kappa^{(k)}_{u,j}|^2.$$

Notice that $\max_{j\in[N]}(|l_{j,i_1}|^2\cdots|l_{j,i_{k-2}}|^2) \le \mathcal{G}^{-(k-2)}$. Therefore,

$$\psi_1(\mathbf{C}^{(k)}_e\mathbf{C}^{(k)'}_e) \le \psi_1^2(G^*_N)[\mathcal{G}^{-1}\mathrm{tr}(G_N)]^{k-2}|\kappa^{(k)}_u|^2.$$

Now for $k \ge 3$, we have $\sigma_1(\mathbf{C}^{(k)}_e) \le \sigma_1(G_N)[\mathcal{G}^{-1}\mathrm{tr}(G_N)]^{\frac{k}{2}-1}|\kappa^{(k)}_u|$. Lemma 2 already shows that $\sigma_j(N^{\frac{(\alpha-1)k}{2}}\Lambda\mathbf{C}^{(k)}_f(\Lambda'^{\otimes(k-1)})) \asymp O(1)$ for $j = 1,2,\ldots,R$. It is sufficient for $\sigma_j(\Lambda\mathbf{C}^{(k)}_f\Lambda'^{\otimes(k-1)}) \gg$

$\sigma_1(\mathbf{C}_e^{(k)})$ if we have

$$N^{\frac{(\alpha-1)k}{2}}\sigma_1(G_N)[\mathcal{G}^{-1}\mathrm{tr}(G_N)]^{\frac{k}{2}-1}\kappa_u^{(k)} = o(1). \tag{22}$$

Furthermore, if there exist some identical eigenvalues in G_N, the eigenvalue decomposition and the eigenvector matrix L are not unique. Hence, under such case, we need that (22) holds for any matrix L belonging to the eigenvector space of G_N.

1.6. Proof of Remark 4.3

First, we give the proof of the convergence rate of the HFA estimator following the normalization conditions that $\Lambda'\Lambda/N = \mathbf{I}_R$ and $\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)\prime}$ be diagonal. Thus we have $\|\Lambda\| = O(\sqrt{N})$ and $\|F\| = O_p(\sqrt{N^{-\alpha}T})$.

Lemma 7 implies that $p\lim_{(T,N)\to\infty}\Lambda'\widehat{\Lambda}^{(k)}/N = \mathbf{I}_R$. Combining this with Lemma 8 we have $H^{(k)} \to \mathbf{I}_R$, we can remove the rotation matrix $H^{(k)}$ and obtain

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(k)} - \Lambda\| = O_p(\sqrt{\frac{\mathrm{tr}(G_N)^{k-1}\log N}{N^{(1-\alpha)k}T}}) + O_p(\sqrt{\frac{N^\alpha}{T}}),\\
&\frac{1}{\sqrt{T}}\|\widehat{F}^{(k)} - F\| = O_p(\frac{1}{\sqrt{N}}) + O_p(\frac{1}{\sqrt{TN^{1-\alpha}}}) + O_p(\sqrt{\frac{\mathrm{tr}(G_N)^{k-1}\log N}{N^{(1-\alpha)k+1}T}}).
\end{aligned} \tag{23}$$

We then derive the limit distribution of $\widehat{\lambda}_i^{(k)}$ and $\widehat{f}_t^{(k)}$ under the normalization conditions $\Lambda'\Lambda/N = \mathbf{I}_R$ and $\widetilde{\mathbf{C}}_f^{(k)}\widetilde{\mathbf{C}}_f^{(k)\prime}$ be diagonal. Notice that $H^{(k)} \to \mathbf{I}_R$ and $Q^{(k)} = \mathbf{I}_R$. Notice that $D^{(k)}$ now is the eigenvalues of $\mathbf{C}_f^{(k)}\mathbf{C}_f^{(k)'}$ when $\Lambda'\Lambda/N = \mathbf{I}_R$, namely $\{D^{(k)}\}_{ii} = \sigma_i^2(\mathbf{C}_f^{(k)})$ for $i = 1, 2, \ldots, R$. Therefore, the limit distribution of $\widehat{\lambda}_i^{(k)}$ can be simplified as

$$\sqrt{TN^{-\alpha}}(\widehat{\lambda}_i^{(k)} - \lambda_i) \xrightarrow{d} N(0, (D^{(k)})^{-1}\mathbf{C}_f^{(k)}\Theta_i^{(k)}\mathbf{C}_f^{(k)'}(D^{(k)})^{-1}). \tag{24}$$

For the limit distribution of $\widehat{f}_t^{(k)}$, by Lemmas 7 and 8, it holds that $H^{(k)} \to \mathbf{I}_R$ and $Q^{(k)} \to \mathbf{I}_R$. We then obtain

$$\sqrt{N}(\widehat{f}_t^{(k)} - f_t) \xrightarrow{d} \mathcal{N}(0, \Phi_t), \tag{25}$$

thus the result in Remark 4.3 follows.

2. Asymptotic results of factor-augmented regressions based on HFA factors

In this section, we will give the asymptotic results of the following factor-augmented regression based on HFA factors:

$$y_{t+h} = \beta' f_t + \gamma' W_t + \epsilon_{t+h}. \tag{26}$$

Prior to that, we need to make some assumptions on the factor-augmented regression:

ASSUMPTION E: Factor-augmented regression
(i) Let $z_t = (f_t', W_t')'$, $\mathbb{E}\|z_t\|^8 < \infty$, $\mathbb{E}(\epsilon_{t+h}|y_t, z_t, y_{t-1}, z_{t-1}, \ldots) = 0$ for any $h > 0$, and z_t and ϵ_t be independent of the idiosyncratic errors e_{is} for all i and s; $\frac{1}{T}\sum_{t=1}^{T} z_t z_t' \xrightarrow{p} \Sigma_{zz} > 0$;
(ii) For any $h > 0$, as $T \to \infty$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t \epsilon_{t+h} \xrightarrow{d} \mathcal{N}(0, \Sigma_{zz,\epsilon}),$$

where $\Sigma_{zz,\epsilon} = p\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(\epsilon_{t+h}^2 z_t z_t')$.

We first consider the properties of the least squares estimates $\widehat{\delta}$ when the k-th order HFA estimates of the factors $\widehat{f}_t^{(k)}$ are used as regressors. Define $\widehat{\delta} = (\frac{1}{\sqrt{N^\alpha}}\widehat{\beta}', \widehat{\gamma}')'$ and $\delta = (\frac{1}{\sqrt{N^\alpha}}\beta' H^{(k)}, \gamma')'$, where $H^{(k)}$ is an $R \times R$ rotation and rescaling matrix and $\|H^{(k)}\| = O_p(\sqrt{N^\alpha})$. Subsequently, the following theorem holds:

Theorem 4. *Suppose Assumptions A–E hold. The following results hold for HFA factor-augmented regression:*

(i) *If* $\sqrt{N^{2\alpha-1}}/T \to 0$ *as* $(N,T) \to \infty$ *and* $\mathrm{tr}(G_N) = o(T^{\frac{2}{k-1}} N^{\frac{k+1}{k-1}(1-\alpha)} (\log N)^{-\frac{1}{k-1}})$, *then* $\widehat{\delta}$ *is a consistent estimator for* δ, *i.e.* $\widehat{\delta} \xrightarrow{p} \delta$.

(ii) *Moreover, if* $\mathrm{tr}(G_N) = o\big(T^{\frac{1}{k-1}} N^{\frac{k+1}{k-1}(1-\alpha)} (\log N)^{-\frac{1}{k-1}}\big)$, $\alpha < 1$ *and* $TN^{1-2\alpha} \to \infty$ *as* $(N,T) \to \infty$, *then*

$$\sqrt{T}(\widehat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, \Sigma_\delta), \tag{27}$$

where $\Sigma_\delta = (\mathcal{A}_0^{(k)\prime})^{-1}\Sigma_{zz}^{-1}\Sigma_{zz,\epsilon}\Sigma_{zz}^{-1}(\mathcal{A}_0^{(k)})^{-1}$ *with* $\mathcal{A}_0^{(k)} = \mathrm{diag}(Q^{(k)}, \mathbf{I})$ *being block diagonal and* $Q^{(k)}$ *defined by Theorem 3. A consistent estimator for* Σ_δ, *denoted by* $\widehat{\mathrm{Avar}(\widehat{\delta})}$, *is*

$$\widehat{\mathrm{Avar}(\widehat{\delta})} = (\frac{1}{T}\sum_{t=1}^{T-h} \widehat{z}_t^{(k)} \widehat{z}_t^{(k)\prime})^{-1} (\frac{1}{T}\sum_{t=1}^{T-h} \widehat{\epsilon}_{t+h}^2 \widehat{z}_t^{(k)} \widehat{z}_t^{(k)\prime}) (\frac{1}{T}\sum_{t=1}^{T-h} \widehat{z}_t^{(k)} \widehat{z}_t^{(k)\prime})^{-1}. \tag{28}$$

Theorem 4 establishes asymptotic normality of factor-augmented regression parameter δ with weak factor models. Using the HFA factors ensures that δ is consistent with asymptotic normality under mild conditions on $\mathrm{tr}(G_N)$ for factor loading strength $\alpha \in [0, 1)$. The case $\alpha = 1$ is invalid since $\widehat{F}^{(k)}$ is inconsistent after rescaled by $\sqrt{N^\alpha}$. Bai & Ng (2006) establish the rate of convergence

and the limiting distribution of δ in a classical strong factor model ($\alpha = 0$). However, their theorems are infeasible for weak factor models.

Suppose the object of interest is the conditional mean of (26). A feasible way is to construct a confidence interval for the conditional mean. Note that the equation

$$\widehat{y}_{T+h|T} - y_{T+h|T} = (\widehat{\delta} - \delta)'\widehat{z}_T^{(k)} + \beta' H^{(k)}(\widehat{f}_T^{(k)} - (H^{(k)})^{-1} f_T) \tag{29}$$

has two components, which arise from estimating δ and f_t. Theorems 3 and 4 show that $\sqrt{N}(\widehat{f}_t^{(k)} - (H^{(k)})^{-1} f_t)$ and $\sqrt{T}(\widehat{\delta} - \delta)$ exhibit asymptotic normality for each t, respectively. Therefore, the following result holds for $\widehat{y}_{T+h|T}$:

Theorem 5. *Let $\widehat{y}_{T+h|T} = \widehat{\delta}'\widehat{z}_T^{(k)}$. Under the conditions of Theorem 4 (ii), then*

$$\frac{(\widehat{y}_{T+h|T} - y_{T+h|T})}{\sqrt{\text{var}(\widehat{y}_{T+h|T})}} \xrightarrow{d} \mathcal{N}(0,1), \tag{30}$$

where $\text{var}(\widehat{y}_{T+h|T}) = \frac{1}{T}\widehat{z}_T^{(k)'}\text{Avar}(\widehat{\delta})\widehat{z}_T^{(k)} + \frac{1}{N}\widehat{\beta}'\text{Avar}(\widehat{f}_T^{(k)})\widehat{\beta}$.

Theorem 5 establishes asymptotic normality of forecasts $\widehat{y}_{T+h|T}$ with HFA factors. Notably, $\widehat{y}_{T+h|T}$ ensure consistency for all $\alpha \in [0,1)$, $TN^{1-2\alpha} \to \infty$ as $(N,T) \to \infty$ and $\text{tr}(G_N)$ satisfies mild condition. As the two terms in $\text{var}(\widehat{y}_{T+h|T})$ vanish at different rates, notice that $\|\widehat{\beta}\| = O_p(\sqrt{N^\alpha})$, and the overall convergence rate for $\widehat{y}_{T+h|T}$ is $\min(\sqrt{N^{1-\alpha}}, \sqrt{T})$. When weak factors are used ($\alpha > 0$), the convergence rate of $\widehat{y}_{T+h|T}$ is slower than $\min(\sqrt{N}, \sqrt{T})$.

Furthermore, when the objective is to forecast $y_{T+h} = y_{T+h|T} + \epsilon_{T+h}$, the forecasting error $\widehat{\epsilon}_{T+h} = \widehat{y}_{T+h|T} - y_{T+h} = (\widehat{y}_{T+h|T} - y_{T+h|T}) - \epsilon_{T+h}$. Hence, if $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$, we have $\widehat{\epsilon}_{T+h} \sim \mathcal{N}(0, \sigma_\epsilon^2 + \text{var}(\widehat{y}_{T+h|T}))$, a consistent estimate of σ_ϵ^2 is $\frac{1}{T}\sum_{t=1}^{T}\widehat{\epsilon}_t^2$. Once estimators for $\widehat{\text{var}(\widehat{y}_{T+h|T})}$ are given, prediction intervals can be easily constructed. For example, the 95% confidence interval for the conditional mean $y_{T+h|T}$ is

$$\left(\widehat{y}_{T+h|T} - 1.96\sqrt{\widehat{\text{var}(\widehat{y}_{T+h|T})}}, \widehat{y}_{T+h|T} + 1.96\sqrt{\widehat{\text{var}(\widehat{y}_{T+h|T})}}\right)$$

and the 95% confidence interval for the forecasting variable y_{T+h} is

$$\left(\widehat{y}_{T+h|T} - 1.96\sqrt{\widehat{\sigma}_\epsilon^2 + \widehat{\text{var}(\widehat{y}_{T+h|T})}}, \widehat{y}_{T+h|T} + 1.96\sqrt{\widehat{\sigma}_\epsilon^2 + \widehat{\text{var}(\widehat{y}_{T+h|T})}}\right).$$

A consistent estimate of $\text{var}(\widehat{y}_{T+h|T})$ is given by

$$\widehat{\text{var}(\widehat{y}_{T+h|T})} = \frac{1}{T}\widehat{z}_T^{(k)\prime}\widehat{\text{Avar}(\widehat{\delta})}\widehat{z}_T^{(k)} + \frac{1}{N}\widehat{\beta}'\widehat{\text{Avar}(\widehat{f}_T^{(k)})}\widehat{\beta}, \tag{31}$$

where $\widehat{\text{Avar}(\widehat{\delta})}$ is given in Theorem 4 and $\widehat{\text{Avar}(\widehat{f}_T^{(k)})} = \widehat{\Phi}_T^{(k)}$. Following Bai & Ng (2006), the $R \times R$ matrix $\widehat{\Phi}_t^{(k)}$ can be

$$\widehat{\Phi}_t^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \widehat{e}_{it}^{(k)2}\widehat{\lambda}_i^{(k)}\widehat{\lambda}_i^{(k)\prime}, \tag{32}$$

where $\widehat{e}_{it}^{(k)} = x_{it} - \widehat{\lambda}_i^{(k)\prime}\widehat{f}_t^{(k)}$. Notice that $\|\widehat{\lambda}_i^{(k)}\| = O_p(\sqrt{N})$ and $\|\lambda_i\| = O_p(\sqrt{N^{1-\alpha}})$, so we have

$$\widehat{\Phi}_t^{(k)} = \frac{1}{N^{1-\alpha}}\sum_{i=1}^{N} \widehat{e}_{it}^{(k)2}(\widehat{\lambda}_i^{(k)}/\sqrt{N^{\alpha}})(\widehat{\lambda}_i^{(k)}/\sqrt{N^{\alpha}})'. \tag{33}$$

As all the elements in (33) are consistent estimators, $\widehat{\Phi}_t^{(k)}$ is a consistent estimate of Φ_t.

2.1. Proof of Theorem 4

Let $z_t = (f_t', W_t')'$, $\widehat{z}_t^{(k)} = (\sqrt{N^{\alpha}}\widehat{f}_t^{(k)\prime}, W_t')'$, and $H^{(k)} = \widetilde{\mathbf{C}}_f^{(k)}(\frac{\Lambda'\Lambda}{N^{1-\alpha}})^{\otimes(k-1)}\widetilde{\mathbf{C}}_f^{(k)\prime}(\frac{\Lambda'\widehat{\Lambda}^{(k)}}{N^{1-\alpha}})(\widetilde{D}_{NT}^{(k)})^{-1}$. We directly start to prove the asymptotic distribution of $\widehat{\delta}$, the consistency of $\widehat{\delta}$ can be easily derived from it, hence we omit it here. Adding and subtracting terms, the regression model can be written as

$$\begin{aligned} y_{t+h} &= \beta' f_t + \gamma' W_t + \epsilon_{t+h} \\ &= \beta' H^{(k)}\widehat{f}_t^{(k)} + \gamma' W_t + \epsilon_{t+h} + \beta' H^{(k)}\{(H^{(k)})^{-1}f_t - \widehat{f}_t^{(k)}\} \\ &= \delta'\widehat{z}_t^{(k)} + \epsilon_{t+h} + \beta' H^{(k)}\{(H^{(k)})^{-1}f_t - \widehat{f}_t^{(k)}\}. \end{aligned}$$

In matrix notation, $Y = \widehat{z}^{(k)}\delta + \epsilon + \{F(H^{(k)\prime})^{-1} - \widehat{F}^{(k)}\}H^{(k)}\beta$, where $Y = (y_{h+1}, \ldots, y_T)'$, $\epsilon = (\epsilon_{h+1}, \ldots, \epsilon_T)'$, and $\widehat{z}^{(k)} = (\widehat{z}_1^{(k)}, \ldots, \widehat{z}_{T-h}^{(k)})'$. The ordinary least squares estimator is $\widehat{\delta} = (\widehat{z}^{(k)\prime}\widehat{z}^{(k)})^{-1}\widehat{z}^{(k)\prime}Y$. Thus,

$$\sqrt{T}(\widehat{\delta} - \delta) = (T^{-1}\widehat{z}^{(k)\prime}\widehat{z}^{(k)})^{-1}\frac{1}{\sqrt{T}}\widehat{z}^{(k)\prime}\epsilon + (T^{-1}\widehat{z}^{(k)\prime}\widehat{z}^{(k)})^{-1}\frac{1}{\sqrt{T}}\widehat{z}^{(k)\prime}\{F(H^{(k)\prime})^{-1} - \widehat{F}^{(k)}\}H^{(k)}\beta. \tag{34}$$

First, notice that $\|\frac{1}{\sqrt{T}}\widehat{z}^{(k)\prime}\{F(H^{(k)\prime})^{-1} - \widehat{F}^{(k)}\}H^{(k)}\beta\| \leq \|\frac{1}{\sqrt{T}}\widehat{z}^{(k)\prime}\{F(H^{(k)\prime})^{-1} - \widehat{F}^{(k)}\}\|\|H^{(k)}\|O(1)$ $= O_p(\frac{1}{\sqrt{N^{1-\alpha}}}) + O_p(\frac{1}{\sqrt{TN^{1-2\alpha}}}) + O_p(\sqrt{\frac{\text{tr}(G_N)^{k-1}N^{(\alpha-1)k}\log N}{N^{1-\alpha}T}})$ by Theorem 2 and $\|H^{(k)}\| = O_p(\sqrt{N^{\alpha}})$. Therefore, the second term on the right of (34) is $o_p(1)$ if $\text{tr}(G_N) = o\big(T^{\frac{1}{k-1}}N^{\frac{k+1}{k-1}(1-\alpha)}(\log N)^{-\frac{1}{k-1}}\big)$, $TN^{1-2\alpha} \to \infty$ and $\alpha < 1$. For the first term on the right of (34), $\frac{1}{\sqrt{T}}\widehat{z}^{(k)\prime}\epsilon = \frac{1}{\sqrt{T}}(\sqrt{N^{\alpha}}\epsilon'\widehat{F}^{(k)}, \epsilon'W)'$.

Now,

$$\sqrt{\frac{N^\alpha}{T}}\widehat{F}^{(k)\prime}\epsilon = (H^{(k)}/\sqrt{N^\alpha})^{-1}\frac{1}{\sqrt{T}}F'\epsilon + \sqrt{\frac{N^\alpha}{T}}(\widehat{F}^{(k)} - F(H^{(k)\prime})^{-1})'\epsilon,$$

where the second term $\sqrt{\frac{N^\alpha}{T}}(\widehat{F}^{(k)} - F(H^{(k)\prime})^{-1})'\epsilon = o_p(1)$, which provided that the above conditions. Thus, $\frac{1}{\sqrt{T}}\widehat{z}^{(k)\prime}\epsilon = \frac{1}{\sqrt{T}}(\epsilon' F(H^{(k)\prime}/\sqrt{N^\alpha})^{-1}, \epsilon' W)' + o_p(1) = \frac{1}{\sqrt{T}}\mathcal{A}z'\epsilon + o_p(1)$, where $\mathcal{A}^{(k)} = \text{diag}\big((H^{(k)}/\sqrt{N^\alpha})^{-1}, \mathbf{I}\big)$ is a block diagonal matrix. Thus,

$$\sqrt{T}(\widehat{\delta} - \delta) = (T^{-1}\widehat{z}^{(k)\prime}\widehat{z}^{(k)})^{-1}\frac{1}{\sqrt{T}}z'\epsilon + o_p(1).$$

Since $\frac{1}{\sqrt{T}}z'\epsilon \xrightarrow{d} \mathcal{N}(0, \Sigma_{zz,\epsilon})$ by Assumption E(ii), the above is asymptotically normal. As $(\frac{H^{(k)}}{\sqrt{N^\alpha}})^{-1} \xrightarrow{p} Q^{(k)}$ by Lemmas 7 and 8. Define $\mathcal{A}_0^{(k)} = \text{diag}(Q^{(k)}, \mathbf{I})$; now, $T^{-1}\widehat{z}^{(k)\prime}\widehat{z}^{(k)} = \mathcal{A}^{(k)}(T^{-1}z'z)\mathcal{A}^{(k)\prime} + o_p(1) \xrightarrow{p} \mathcal{A}_0^{(k)}\Sigma_{zz}\mathcal{A}_0^{(k)\prime}$. The limiting variance is

$$\begin{aligned}\Sigma_\delta =& (\mathcal{A}_0^{(k)}\Sigma_{zz}\mathcal{A}_0^{(k)\prime})^{-1}(\mathcal{A}_0^{(k)}\Sigma_{zz,\epsilon}\mathcal{A}_0^{(k)\prime})(\mathcal{A}_0^{(k)}\Sigma_{zz}\mathcal{A}_0^{(k)\prime})^{-1} \\ =& (\mathcal{A}_0^{(k)\prime})^{-1}\Sigma_{zz}^{-1}\Sigma_{zz,\epsilon}\Sigma_{zz}^{-1}(\mathcal{A}_0^{(k)})^{-1}.\end{aligned}$$

As $f_t = H^{(k)}\widehat{f}_t^{(k)} + o_p(1)$ and $z_t = (f_t', W_t')'$, we have $\mathcal{A}^{(k)}(\frac{1}{T}\sum_{t=1}^T \epsilon_{t+h}^2 z_t z_t')\mathcal{A}^{(k)\prime} = \frac{1}{T}\sum_{t=1}^T \widehat{\epsilon}_{t+h}^2 \widehat{z}_t^{(k)}\widehat{z}_t^{(k)\prime} + o_p(1)$. Therefore, $\widehat{\text{Avar}(\widehat{\delta})} = (\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t^{(k)}\widehat{z}_t^{(k)\prime})^{-1}(\frac{1}{T}\sum_{t=1}^{T-h}\widehat{\epsilon}_{t+h}^2\widehat{z}_t^{(k)}\widehat{z}_t^{(k)\prime})(\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t^{(k)}\widehat{z}_t^{(k)\prime})^{-1}$ is a consistent estimator for Σ_δ.

2.2. Proof of Theorem 5

We first rewrite the prediction error

$$\begin{aligned}\widehat{y}_{T+h|T} - y_{T+h|T} =& \widehat{\beta}'\widehat{f}_T^{(k)} + \widehat{\gamma}'W_T - \beta' f_T - \gamma' W_T \\ =& \frac{1}{\sqrt{N^\alpha}}(\widehat{\beta} - H^{(k)\prime}\beta)'\sqrt{N^\alpha}\widehat{f}_T^{(k)} + \beta' H^{(k)}(\widehat{f}_T^{(k)} - (H^{(k)})^{-1}f_T) + (\widehat{\gamma} - \gamma)'W_T \\ =& \frac{1}{\sqrt{T}}\widehat{z}_T^{(k)\prime}\{\sqrt{T}(\widehat{\delta} - \delta)\} + \frac{1}{\sqrt{N}}\beta' H^{(k)}\{\sqrt{N}(\widehat{f}_T^{(k)} - (H^{(k)})^{-1}f_T)\}.\end{aligned}$$

Both $\sqrt{T}(\widehat{\delta} - \delta)$ and $\sqrt{N}(\widehat{f}_T^{(k)} - (H^{(k)})^{-1}f_T)$ are asymptotically normal. They are also asymptotically independent because the limit of $\sqrt{T}(\widehat{\delta} - \delta)$ is determined by $(\epsilon_1, \epsilon_2, \ldots, \epsilon_T)$, and that of $\sqrt{N}(\widehat{f}_T^{(k)} - (H^{(k)})^{-1}f_T)$ is determined by e_{iT} for $i = 1, 2, \ldots, N$. Noting that $T^{-1/2}(\widehat{z}_T - z_T) = o_p(1)$, an estimate for the variance of $\frac{1}{\sqrt{T}}z_T'\{\sqrt{T}(\widehat{\delta} - \delta)\}$ is $\frac{1}{T}z_T'\text{Avar}(\widehat{\delta})z_T$, which, in turn, is estimated by $\frac{1}{T}\widehat{z}_T^{(k)\prime}\text{Avar}(\widehat{\delta})\widehat{z}_T^{(k)}$. Similarly, an estimate for the variance of the second term is $\frac{1}{N}(\beta' H^{(k)})'\text{Avar}(\widehat{f}_T^{(k)})(\beta' H^{(k)})$, which, in turn, is estimated by $\frac{1}{N}\widehat{\beta}'\text{Avar}(\widehat{f}_T^{(k)})\widehat{\beta}$. Thus, $(\widehat{y}_{T+h|T} - y_{T+h|T})/\text{var}(\widehat{y}_{T+h|T})^{1/2} \xrightarrow{d} \mathcal{N}(0, 1)$.

3. HFA estimates in presence of Gaussian factors

In the main paper, we assume that all common factors are non-Gaussian. In this section, we extend the approach to the case of both Gaussian and non-Gassuian factors. In subsection 4.1 we describe the approach for estimating the number of factors. In subsection 4.2, we provide an iterative estimation approach of Gaussian and non-Gaussian factors in the presence of independent Gaussian factors.

3.1. Two-step estimation of the number of factors

In the main paper Lu et al. (2024), we show that the number of non-Gaussian factors can be estimated consistently by maximizing the GER criterion. If some factors are Gaussian and independent with the non-Gaussian factors, the GER estimator is not consistent for the number of factors R, because Assumption A(iii) is violated ($\mathbf{C}_f^{(k)}$ is not full rank). The estimation of Gaussian and non-Gassuian factors in the observed factor model has been studied in Lu & Huang (2022). In this subsection, we propose a two-step estimation to solve this practical problem. To be specific, we first use the GER estimator to select the number of non-Gaussian factors. Given a consistent estimate of the number of non-Gaussian factors, we then estimate the remaining Gaussian factors based on a filtered series without the non-Gaussian factor structure by using the estimator of Ahn & Horenstein (2013). The two-step estimation guarantees the consistency of factor number estimation when the Gaussian factors exist. Before we introduce the procedure, it is necessary to redefine the factor model. We denote the non-Gaussian factors as f_{ht} and the Gaussian factors as f_{gt}, the corresponding factor loadings are Λ_h and Λ_g, respectively. Therefore, the factor model can be rewritten as

$$X = F_h\Lambda_h' + F_g\Lambda_g' + E = (F_h, F_g)\begin{pmatrix}\Lambda_h' \\ \Lambda_g'\end{pmatrix} + E = F\Lambda' + E, \tag{35}$$

where $F = (F_h, F_g)$ is the factor matrix and $\Lambda = (\Lambda_h, \Lambda_g)$ is the factor loading matrix. Non-Gaussian and Gaussian factors are assumed to be independent. Let R_h and R_g be the number of non-Gaussian and Gaussian factors, respectively. When $R_g = 0$, the GER estimator give a consistent estimation of the number of factors, namely $\widehat{R}_{GER} = \widehat{R}_{h,GER} \to R$. When $R_g > 0$, the GER estimator only give a consistent estimation of R_h. Therefore, we shall assume a known R_h and give the consistent estimates of the number of Gaussian factors R_g using the filtered series:

$$\omega_{g,t} \equiv x_t - \Lambda_h f_{ht} = \Lambda_g f_{gt} + e_t. \tag{36}$$

Since the series $\omega_{g,t}$ only contains the Gaussian factor structure, it is a traditional factor number selection problem in the mean-variance framework. Several authors have proposed effective methods to estimate R_g, e.g. Bai & Ng (2002) and Onatski (2010). We recommend to use the Ahn & Horenstein (2013)'s Eigenvalue Ratio test to estimate the number of Gaussian factors R_g. Since $\omega_{g,t}$ is unobserved, R_g is estimated by the consistent estimates $\widehat{\omega}_{g,t} = x_t - \widehat{\Lambda}_h \widehat{f}_{ht}$. In the next section, we will discuss the factor extraction methods used to estimate Λ_h and f_{ht} consistently. The consistency of R_g is not affected when the number of non-Gaussian factors R_h, the non-Gaussian factors f_{ht}, and corresponding factor loadings Λ_h, are unknown or consistently estimated. This is because $P(\widehat{R}_g = R_g) = P(\widehat{R}_g = R_g, \widehat{R}_h = R_h) + P(\widehat{R}_g = R_g, \widehat{R}_h \neq R_h)$, and $P(\widehat{R}_g = R_g, \widehat{R}_h \neq R_h) \leq P(\widehat{R}_h \neq R_h) = o(1)$. Thus

$$\begin{aligned} P(\widehat{R}_g = R_g) &= P(\widehat{R}_g = R_g, \widehat{R}_h = R_h) + o(1), \\ &= P(\widehat{R}_g = R_g | \widehat{R}_h = R_h) P(\widehat{R}_h = R_h) + o(1), \\ &= P(\widehat{R}_g = R_g | \widehat{R}_h = R_h) + o(1), \end{aligned} \tag{37}$$

provided that $P(\widehat{R}_h = R_h) \to 1$. In addition, denote $\widehat{R}_g^*$ as the number of Gaussian factors estimated by the true series $\omega_{g,t}$. Similarly, we have $P(\widehat{R}_g = R_g) = P(\widehat{R}_g = R_g | \widehat{R}_h = R_h, \widehat{R}_g = \widehat{R}_g^*) + o(1)$ because $P(\widehat{R}_g^* = R_g) \to 1$ and $P(\widehat{R}_g = \widehat{R}_g^*) \to 1$ ($\widehat{\omega}_{g,t} \to \omega_{g,t}$). In summary, given the consistent estimates of R_h, F_h and Λ_h, the consistency of R_g can be guaranteed based on $\widehat{\omega}_{g,t}$.

Overall, by giving consistent estimates of R_h and R_g respectively, we can get a consistent estimate of the number of factors $\widehat{R} = \widehat{R}_h + \widehat{R}_g$. It should be noticed that for all $\alpha \in [0, 1]$, the GER estimator is consistent if the conditions in Theorem 1 are satisfied, however, consistency of $\widehat{R}_g$ only holds in $\alpha \in [0, 0.5)$, see in Freyaldenhoven (2022). For some Gaussian factors with weak explanatory power ($\alpha > 0.5$), the existing covariance-based approaches cannot identify them successfully, this can be improved by using the GER estimator if factors have information on their higher-order moments.

3.2. Iterative estimation of factors and loadings

In this subsection, we present the procedure used to estimate factors and factor loading in the presence of independent Gaussian factors. The notations in Section 4.1 continue to be used in this section. We assume that the number of non-Gaussian factors R_h and the number of Gaussian factors R_g is known (or estimated consistently). We provide a two-step procedure to estimate factors and loadings.

First, ignore (F_g, Λ_g) and given R_h, we can define the error terms $e_t^* = \Lambda_g f_{gt} + e_t$, the original factor model (35) reduces to the structure $x_t = \Lambda_h f_{ht} + e_t^*$. Following the identification condition

in the main paper, Λ_h can be estimated by

$$\widehat{\Lambda}_h^{(k)} = \arg\max_{\Lambda_h} \mathrm{tr}\big\{\Lambda_h'(\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime})\Lambda_h\big\}, \tag{38}$$

subject to the constraint $\frac{1}{N}\Lambda_h'\Lambda_h = \mathbf{I}_{R_h}$. The HFA estimate of Λ_h subject to the constraint, denotes that $\widehat{\Lambda}_h^{(k)}$, is $\sqrt{N}$ times the eigenvectors corresponding to the R_h largest eigenvalues of the $N \times N$ matrix $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime}$. Given $\widehat{\Lambda}_h^{(k)}$, the non-Gaussian factors can be obtained as $\widehat{F}_h^{(k)} = X\widehat{\Lambda}_h^{(k)}/N$.

Next, given $(\widehat{F}_h^{(k)}, \widehat{\Lambda}_h^{(k)})$ and R_g, we define the matrix $\Omega_g = (\omega_{g,1}, \omega_{g,2}, \ldots, \omega_{g,T})' \in \mathbb{R}^{T\times N}$ with $\omega_{g,t} = x_t - \widehat{\Lambda}_h^{(k)}\widehat{f}_{ht}^{(k)}$. Then, based on a similar argument, the original model (35) reduces to the structure $\omega_{g,t} = \Lambda_g f_{gt} + e_t$. Therefore, estimates of Λ_g can be obtained by maximizing the objective function:

$$\widehat{\Lambda}_g = \arg\max_{\Lambda_g} \mathrm{tr}\big\{T^{-1}\Lambda_g'\Omega_g'\Omega_g\Lambda_g\big\}, \tag{39}$$

subject to the constraint $\frac{1}{N}\Lambda_g'\Lambda_g = \mathbf{I}_{R_g}$. It is an eigenvalue decomposition of matrix $\frac{1}{NT}\Omega_g'\Omega_g$. The estimate of Λ_g, denoted as $\widehat{\Lambda}_g$, is $\sqrt{N}$ times the eigenvectors corresponding to the R_g largest eigenvalues of the matrix $\frac{1}{T}\Omega_g'\Omega_g$ and the estimated Gaussian factors $\widehat{F}_g = \Omega_g\widehat{\Lambda}_g/N$.

Moreover, following Ando & Bai (2016)'s alternating regressions, the estimators can be obtained by using the following iterative algorithm:

Step 1 Initialize the number of non-Gaussian factors (R_h) by the GER estimator or priori knowledge, and the non-Gaussian factors and the corresponding factor-loading matrix $\{F_{h,0}, \Lambda_{h,0}\}$ by solving (38).

Step 2 Given $\{\widehat{F}_{h,0}^{(k)}, \widehat{\Lambda}_{h,0}^{(k)}\}$, initialize the number of Gaussian factors (R_g) by Ahn & Horenstein (2013)'s ER estimator on the filtered series $\omega_{g,t} = x_t - \widehat{\Lambda}_h^{(k)}\widehat{f}_{ht}^{(k)}$. Then, initialize the Gaussian factors and the corresponding factor-loading matrix $\{F_{g,0}, \Lambda_{g,0}\}$ by solving (39).

Step 3 Given $\{\widehat{F}_{g,0}, \widehat{\Lambda}_{g,0}\}$, updating the non-Gaussian factors and the corresponding factor-loading matrix $\{\widehat{F}_{h,1}^{(k)}, \widehat{\Lambda}_{h,1}^{(k)}\}$ by solving (38).

Step 4 Repeat Step 2 and Step 3 until convergence is achieved.

Remark 3.1.

- *The non-Gaussian factors and corresponding factor loadings* $\{F_h, \Lambda_h\}$ *initialized by optimizing* (38) *with the error terms* $e_t^* = \Lambda_g f_{gt} + e_t$. *Notice that* e_t^* *satisfies Assumption C. Hence* $\{\widehat{F}_h^{(k)}, \widehat{\Lambda}_h^{(k)}\}$ *are consistent estimations by Theorem 2. Furthermore, using the filtered series* $\widehat{\omega}_{g,t} = x_t - \widehat{\Lambda}_h^{(k)}\widehat{f}_{ht}^{(k)}$ *one can then initialize* R_g *and the Gaussian factors. This two-step*

estimation preserves consistency of the estimators since the presence of Gaussian factors doesn't affect the asymptotic properties of the proposed estimator of non-Gaussian factors by Assumption A(iii).

4. HFA estimates with non-stationary data

Inference on factors using HFA requires the k-th order cumulant of the factors f_t to converge to a constant matrix as $T \to \infty$, and the sample cumulant converges to its population counterpart, see in Assumption A(ii). This assumption is clearly validated in case of stationary data. The goal of this section is to explore how HFA can be applied in case of non-stationary data.

(a) Time trend

A first case of non-stationarity occurs when the data has a deterministic time trend. In this case, we recommend to transform the data to stationarity before applying HFA. To illustrate this, let us consider a single-factor model with a time trend as follows:

$$x_{it} = \mu_i + \beta_i t + \lambda_i f_t + \epsilon_{it}, \quad i = 1, \ldots, N;\ t = 1, \ldots, T. \tag{40}$$

where μ_i is fixed effect, λ_i and f_t are factor loading and non-Gaussian factor respectively, ϵ_{it} is normal idiosyncratic error. Without loss of generality, we assume $\mathbb{E}(f_t) = 0$, $\mathbb{E}(\epsilon_{it}) = 0$, and f_t and ϵ_{it} are mutually independent. The point is whether HFA can estimate f_t consistently, especially under the weak factor models.

We first show that performing the HFA directly on x_{it} fails to obtain a consistent estimator of f_t, even in a strong factor model. Moreover, if T is large enough, we cannot identify the underlying factor structure based on the multi-cumulant of x_{it}. Without any prior knowledge about the non-stationary of x_{it}, we will detect two factors by the eigenvalue of $\widetilde{\mathbf{C}}_x^{(k)}$. This is because the time trend can be regarded as a factor with a given finite sample size (N, T). Then the factor model can be rewritten as

$$\bar{x}_{it} = \beta_i f_t^* + \lambda_i f_t + \epsilon_{it},$$

where $\bar{x}_{it} = x_{it} - T^{-1}\sum_{t=1}^{T} x_{it}$, $f_t^* = t - \frac{T+1}{2}$ with $T^{-1}\sum_{t=1}^{T} f_t^* = 0$ for a given T, and $T^{-1}\sum_{t=1}^{T}(f_t^*)^2 = (T+2)(T+4)/12$. With an appropriate size of T, one may regard f_t^* as a strong explanatory factor as the scree plot shows there are two dominant eigenvalues. Therefore, we can obtain HFA

estimates of "the two factors" with rotation matrix $H = [h_{ij}]$ such as

$$\widehat{f}_{1t} \approx h_{11} f_t^* + h_{12} f_t = -h_{11}(T+1)/2 + h_{11} t + h_{12} f_t,$$
$$\widehat{f}_{2t} \approx h_{21} f_t^* + h_{22} f_t = -h_{21}(T+1)/2 + h_{21} t + h_{22} f_t.$$

The estimated two factors are a linear combination of the underlying factor f_t and time trending, hence the estimator is non-stationary. This indicates the HFA estimator cannot recover f_t by using the non-stationary data x_{it}. Moreover, when $T \to \infty$, the eigenvalue provided by $\beta_i f_t^*$ will dominate the factor structure $\lambda_i f_t + \epsilon_{it}$, hence only the time trends can be detected by the GER test if we based on the original data x_{it}.

A solution is to conduct HFA on the first-order difference on x_{it} for each i and thus transforming them into stationary process. This is also suggested by McCracken & Ng (2016) to deal with the variables in the FRED-MD dataset, and we follow their proposed transformation rules in our application. The model in first-differenced form is

$$\Delta x_{it} = \beta_i + \lambda_i \Delta f_t + \Delta \epsilon_{it} \tag{41}$$

for $t = 2, 3, \ldots, T$ and $i = 1, 2, \ldots, N$. If λ_i, Δf_t and $\Delta \epsilon_{it}$ satisfies Assumption A-C in the main paper. Then under mild conditions, the HFA estimator is consistent such that $\widehat{\Delta f_t} = \Delta f_t + o_p(1)$. To recover f_t, we let $\widehat{f}_t = \sum_{s=2}^{t} \widehat{\Delta f_s}$, we use the following simulation to confirm that $\widehat{f}_t = f_t - f_1 + o_p(1)$, which implies $\widehat{f}_t$ is uniformly consistent for f_t (up to a shift factor f_1), see e.g. **?**. In addition, using Δf_t directly in factor-augmented regression is also common in literature, see e.g. Bai & Ng (2013), McCracken & Ng (2016).

Numeric evidence of the accuracy of HFA when using first-order differenced data generated by the model with time trend in (40) is provided in Figure 4. We consider (40) with non-Gaussian factor f_t ($\eta_1 = 0.5, p_1 = 1, q_1 = \infty$; skewness $= 1.244$, ex-kurtosis $= 4.920$). u_t is normally distributed. $\beta_i \sim \mathcal{N}(0, 0.01)$ and $\mu_i \sim \mathcal{N}(0, 0.01)$. $\sigma_j(G_N) = j^{-0.544}$ for $j \in [N]$. The sample size is set as $N = \{50, 100, 300\}$, $T = \{500, 1000\}$. The factor strength is set as $\alpha = \{0, 0.25, 0.5, 0.75, 1\}$. We study the estimation error of estimated differenced factor $\widehat{\Delta f}_t$ and original factor $\widehat{f}_t$. We use the maximum square error $\max_{t=2,\ldots,T}(\widehat{\Delta f}_t - \Delta f_t)^2$ and $\max_{t=2,\ldots,T}(\widehat{f}_t - f_t)^2$ to measure the accuracy of the estimators. Figure 4 provide simulation evidence of the consistency of the HFA estimators for both Δf_t and f_t in non-stationary data with a time trend. The estimation error decreases as (N, T) increases for all factor strength α. This result strongly supports that differencing x_{it} into stationary before HFA can avoid the effect of the potential time trend in x_{it}.

$\sim$ Insert Figure 4 Here $\sim$

(b) Unit root processes

A second case of non-stationarity is when the factors or (and) idiosyncratic errors are unit root processes, i.e. I(1) processes. To illustrate this, let us consider an I(1) single-factor model as follows:

$$x_{it} = \mu_i + \lambda_i f_t + \epsilon_{it}, \quad f_t \sim \mathrm{I}(1), \quad \epsilon_{it} \sim \mathrm{I}(1), \tag{42}$$

where μ_i is fixed effect. Without loss of generality, we assume Δf_t and Δe_{it} satisfies the assumptions in the main paper. Intuitively, the model (44) in first-differenced form is

$$\Delta x_{it} = \lambda_i \Delta f_t + \Delta \epsilon_{it}. \tag{43}$$

Notice that Δf_t and $\Delta \epsilon_{it}$ are stationary and satisfies the assumptions of HFA in the differenced model (43). Thus the consistent estimates of Δf_t can be obtained. However, establishing the corresponding theoretical properties requires further research, which also includes unit root tests for latent factors and idiosyncratic errors. We leave it here and take it as a future research topic. Numeric evidence of the accuracy of HFA when using first-order differenced data generated by the I(1) model in (44) is provided in Figure 5. Δf_t and Δe_{it} are generated as the same in (40). We still use the maximum square error $\max_{t=2,\ldots,T}(\widehat{\Delta f}_t - \Delta f_t)^2$ and $\max_{t=2,\ldots,T}(\widehat{f}_t - f_t)^2$ to measure the accuracy of the estimators. The estimation error of both $\widehat{\Delta f}_t$ and $\widehat{f}_t$ decreases as N increases for all factor strength α. This result also supports that differencing x_{it} into stationary before HFA can avoid the effect of latent unit root process.

$\sim$ Insert Figure 5 Here $\sim$

(c) Structural break

A third case of non-stationarity is when the factor process has a structural break at a fixed point. To illustrate this, let us consider a single-factor model as follows:

$$\begin{aligned} &x_{it} = \lambda_i f_t + \epsilon_{it}, \quad i = 1, \ldots, N;\ t = 1, \ldots, T; \\ &f_t \sim p_1, 1 \le t \le T_0, \quad \text{and} \quad f_t \sim p_2, T_0 < t \le T, \end{aligned} \tag{44}$$

where p_1 and p_2 are different non-Gaussian distributions, T_0 is the fixed change point. After simple calculation, we can get $\mathbf{C}_f^{(k)} \equiv \lim_{T\to\infty}(\frac{T_0}{T}\mathbf{C}_{f,p_1}^{(k)} + \frac{T-T_0}{T}\mathbf{C}_{f,p_2}^{(k)}) = \mathbf{C}_{f,p_2}^{(k)}$, where $\mathbf{C}_{f,p_1}^{(k)}$ and $\mathbf{C}_{f,p_2}^{(k)}$ are the k-th order multi-cumulant of p_1 and p_2, respectively. If $\mathbf{C}_{f,p_2}^{(k)}$ is a nonzero constant, the Assumption A (ii) of the main paper is satisfied with a fixed change point T_0, since the break is on

a small portion and doesn't contribute to the limit. Under this case, HFA can estimate the factor consistently as $(N,T) \to \infty$.

5. Alternative factor extraction approaches

The proposed HFA complements existing methods for factor extraction such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Moment Component Analysis (MCA). We briefly discuss each of them below.

The objective of PCA is to estimate factors and factor loadings through an eigenvalue decomposition of the covariance (or correlation) matrix. The objective function is as follows

$$\begin{aligned} \mathcal{L}_{PCA}(F,\Lambda) &= (NT)^{-1}\sum_{t=1}^{T}\|x_t - \Lambda f_t\|^2, \\ \text{s.t.} \quad & \frac{1}{N}\Lambda'\Lambda = \mathbf{I}_R. \end{aligned} \tag{45}$$

This problem is identical to maximizing $\text{tr}(\Lambda' X' X \Lambda)$. The estimated factor loading matrix, denoted as $\widehat{\Lambda}_{[PCA]}$, is $\sqrt{N}$ times the eigenvectors corresponding to the R largest eigenvalues of the $N \times N$ matrix $X'X$. Given $\widehat{\Lambda}_{[PCA]}$, the factor matrix can be obtained as $\widehat{F}_{[PCA]} = X\widehat{\Lambda}_{[PCA]}/N$. As shown in Remark 4.1 of the main paper, PCA is a special case of HFA when $k = 2$. In PCA, we rotate the covariance matrix of the data and shrink the variance of the data into a very few factors to achieve dimensionality reduction. Compare to PCA, an intuitive interpretation is that HFA rotation the co-skewness matrix (or co-kurtosis matrix) of the data and shrink the skewness or kurtosis of the data into a low number of factors to achieve dimensionality reduction. Since these higher-order multi-cumulants contain less noise caused by the idiosyncratic errors, we can efficiently extract factors with weak explanatory power. We can understand this by Figure 1. Figure 1(a) give the heat map of $\widetilde{\mathbf{C}}_c^{(2)}$, $\widetilde{\mathbf{C}}_e^{(2)}$ and $\widetilde{\mathbf{C}}_x^{(2)}$. One can observe that the heat map structure spanned by the covariance of common component c_{it} is masked by the heat map structure of that of the error terms e_{it}, which leads to the structure of c_{it} cannot be observed clearly in the heat map of $\widetilde{\mathbf{C}}_x^{(2)}$. In contrast, Figure 1(b) shows that the heat map structure spanned by the third-order cumulant of common component c_{it} can be clearly observed in $\widetilde{\mathbf{C}}_x^{(3)}\widetilde{\mathbf{C}}_x^{(3)'}$ after adding noise e_{it}.

ICA is an improved method of PCA designed to extract independent non-Gaussian factors. A classic ICA involves two steps: First, PCA is used to obtain the uncorrelated factors. This step is usually referred to a pre-whitening of the data, and the uncorrelated factors are named pre-whitened data. Let denote the vector of pre-whitened data $z_t = [z_{1t}, z_{2t}, \ldots, z_{Rt}]' = \widehat{V}^{-1/2}\widehat{f}_{t[PCA]}$, where $\widehat{V}$ is $R \times R$ eigenvalue matrix of $\widehat{F}_{[PCA]}$, to satisfy $\frac{1}{T}\sum_{t=1}^{T} z_t z_t' = I_R$. Second, the pre-whitened

data are rotated to minimize their statistical dependence through an objective function. Several objective functions have been proposed to obtain independent factors. A common approach named FastICA (Hyvärinen & Oja,1997) consists of maximizing the negentropy of the pre-whitened data. This method finds an $R \times R$ orthonormal matrix W to maximize the objective function

$$\begin{aligned} \mathcal{L}_{FAST}(W) &= \{\mathbb{E}[G(W'z_t)] - \mathbb{E}[G(v)]\}^2, \\ \text{s.t.} \quad & W'W = \mathbf{I}_R, \end{aligned} \tag{46}$$

where the function G is a proxy of the negentropy and v is the Gaussian variable. For instance, using kurtosis to proxy negentropy ($G(y) = y^4$), one then solves for W to obtain the independent factors $\widehat{F}_{[ICA]} = Z\widehat{W} = \widehat{F}_{[PCA]}\widehat{V}^{-1/2}\widehat{W}$. The corresponding estimated factor loading matrix is $\widehat{\Lambda}_{[ICA]} = \widehat{\Lambda}_{[PCA]}\widehat{V}^{1/2}(\widehat{W}')^{-1}$. It is clear that ICA only changes scale (by $\widehat{V}^{-1/2}$) and rotates direction (by $\widehat{W}$), but the factors are still in the same space spanned by PCA estimators.

Another popular class of ICA method extracts independent factors by performing an eigenvalue decomposition of the higher-order multi-cumulants of the pre-whitened data. The Joint Approximate Diagonalization of Eigenmatrices (JADE) proposed by Cardoso & Souloumiac (1993), which relies on the eigenvalue decomposition of the fourth-order cumulant tensor. This method finds rotation matrix W to minimize the objective function

$$\begin{aligned} \mathcal{L}_{JADE}(W) &= \sum_{iikl \neq ijkl} (\kappa_{W'z_t,ijkl})^2, \\ \text{s.t.} \quad & W'W = \mathbf{I}_R, \end{aligned} \tag{47}$$

where $\kappa_{W'z_t,ijkl}$ is the fourth-order multi-cumulant of independent factors $W'z_t$. It attempts to minimize the sum of off-diagonal squared multi-cumulants which should be zero if $W'z_t$ are independent.

There are two main distinctions between ICA and HFA: First, ICA needs a measure of independence for estimating the non-Gaussian independent factors. In HFA, the singular values of the higher-order multi-cumulants of factors are used to measure non-Gaussianity and estimate the non-Gaussian factors by maximizing the sum of the squared singular values of higher-order multi-cumulants. Therefore, HFA recovers the non-Gaussian factors without considering minimizing statistical independence. When non-Gaussian factors are mutually dependent, ICA cannot estimate the non-Gaussian factors consistently but HFA is still consistent in many cases. Second, ICA cannot improve the linear space spanned by PCA estimators, just up to rotation and scale. In contrast, HFA estimators show more efficiency than PCA estimators in estimating factors and factor loadings.

MCA (Jondeau et al., 2018) is another novel method to extract the factors which drive the higher-order co-moments (such as co-skewness and co-kurtosis) structures. It is based on multilinear eigenvalue decomposition of the higher-order co-moments. We describe the joint-MCA which assumes covariance, co-skewness, and co-kurtosis matrix are driven by the same factors. Denote $\widetilde{\mathbf{M}}_x^{(k)}$ as the sample k-th-order co-moment of response variables and R_m as the number of MCA factors. The objective function of MCA is as follows:

$$\begin{aligned} \mathcal{L}_{MCA}(F,\Lambda) &= \sum_{k=2}^{4} \frac{1}{N^k} \|\widetilde{\mathbf{M}}_x^{(k)} - \Lambda \widetilde{\mathbf{M}}_f^{(k)} (\Lambda'^{\otimes(k-1)})\|^2, \\ \text{s.t.} \quad & \Lambda'\Lambda = I_{R_m}, \end{aligned} \tag{48}$$

where $\widetilde{\mathbf{M}}_f^{(k)}$ is the k-th-order co-moment tensor spanned by MCA factors. MCA attempts to estimate the factors driving the higher-order co-moments, but the asymptotic properties of MCA are unknown.

The Nearest Co-moment (NC) approach proposed by Boudt et al. (2020) estimates the parameters of a latent factor model by minimizing the distance between the sample moments and model-based moments. The objective function of NC is as follows:

$$\mathcal{L}_{NC}(\theta) = (\eta(\theta) - \widehat{\eta}_s)' \widehat{W} (\eta(\theta) - \widehat{\eta}_s), \tag{49}$$

where θ is the parameter vector of the factor model which contains the factor loadings (Λ), moments of factors and errors, $\eta(\theta)$ is the model moments and $\widehat{\eta}_s$ is the sample moments. $\widehat{W}$ is a positive semi-definite weight matrix converging in probability to the positive semi-definite matrix W. The NC approach consistently estimate the factor loading matrix and the moments of factors and errors. However, since the dimension of the weight matrix W increase as $O(N^4)$, this approach is not suitable for the high dimensional case.

6. Alternative approach for selecting the number of non-Gaussian factors

6.1. Generalized Growth Ratio estimator

The main paper presents a generalization of the Eigenvalue Ratio estimator of the number of factors, as introduced by Ahn & Horenstein (2013). The latter also present a growth ratio estimator. Below we provide its generalization to higher order multi-cumulants. The Generalized

Growth Ratio (GGR) function we consider is

$$\begin{aligned}\mathrm{GGR}^{(k)}(r) \equiv & \frac{\ln(V^{(k)}(r-1)/V^{(k)}(r))}{\ln(V^{(k)}(r)/V^{(k)}(r+1))} \\ = & \frac{\ln(1+\widetilde{\mu}_{NT,r}^{(k)*})}{\ln(1+\widetilde{\mu}_{NT,r+1}^{(k)*})}, \quad r=1,2,\ldots,R_{\max},\end{aligned} \tag{50}$$

where $V^{(k)}(r)=\sum_{l=r+1}^{N}\widetilde{\mu}_{NT,l}^{(k)}$ and $\widetilde{\mu}_{NT,r}^{(k)*}=\widetilde{\mu}_{NT,l}^{(k)}/V^{(k)}(r)$. The term "GGR" refers to the Generalized Growth Ratio because as discussed in Remark 3.1 below, it can be approximately interpreted as the growth rate between the two adjacent sums of the residuals. Our proposed GGR estimator for R is the maximizer of $\mathrm{GGR}^{(k)}(r)$:

$$\widehat{R}_{\mathrm{GGR}}^{(k)}=\max_{1\leq r\leq R_{\max}}\mathrm{GGR}^{(k)}(r). \tag{51}$$

Remark 6.1.

- *For the GGR estimator, we have* $V^{(k)}(r)\geq\|\widehat{\mathbf{C}}_{x[r]}^{(k)}-\widetilde{\mathbf{C}}_{x}^{(k)}\|^2$ *(see De Lathauwer et al., 2000), where* $\widehat{\mathbf{C}}_{x[r]}^{(k)}=\widehat{\Lambda}_{[r]}^{(k)}\widehat{\mathbf{C}}_{f[r]}^{(k)}(\widehat{\Lambda}_{[r]}^{(k)\prime}\otimes\ldots\otimes\widehat{\Lambda}_{[r]}^{(k)\prime})$*, are the k-th-order multi-cumulants of the first r factors and factor loadings estimated by Higher-Order EigenValue Decomposition (De Lathauwer et al.,2000). Hence,* $V^{(k)}(r)$ *can approximately represent the residuals between* $\widehat{\mathbf{C}}_{x}^{(k)}(r)$ *and* $\widetilde{\mathbf{C}}_{x}^{(k)}$*.*

- *For the maximum number of non-Gaussian factors* $R_{\max}$*, Ahn & Horenstein (2013) recommend two possible choices for* $R_{\max}$*. First, if we have a priori information about a possible maximum number of factors, for example* $R_{priori,max}$*, we can set* $R_{\max}=2R_{priori,max}$*. This choice is suitable for the case where* $R_{\max}$ *is fixed. Second, consider using a sequence,* $R_{\max}=\min(R_{\max}^{*},0.2[N])$*, where* $R_{\max}^{*}=\#\{r\mid\widetilde{\mu}_{NT,r}^{(2)}\geq\sum_{r=1}^{N}\widetilde{\mu}_{NT,r}^{(2)}/N^2,r\geq 1\}$*. As shown in Lemmas 5 and 6,* $\sum_{r=1}^{N}\widetilde{\mu}_{NT,r}^{(2)}=O_p(N)+O_p(RN^{1-\alpha})$ *and* $\widetilde{\mu}_{NT,r}^{(2)}=O_p(N^{1-\alpha})$ *for* $r=1,2,\ldots,R$*. Thus,* $Prob(R_{\max}^{*}\leq R)\to 0$ *as* $N\to\infty$*.*

Proof of the consistency of GGR estimator: Under the condition $N^{\alpha}/T\to 0$ as $(N,T)\to\infty$ and $\mathrm{tr}(G_N)=o\big(N^{\frac{k}{k-1}(1-\alpha)}T^{\frac{1}{k-1}}(\log N)^{-\frac{1}{k-1}}\big)$, note that $V^{(k)}(R+1)=\sum_{j=R+2}^{N-R}\widetilde{\mu}_{NT,j}^{(k)}$. For each part we have

$$(N-2R-1)\widetilde{\mu}_{NT,N-R}^{(k)}\leq\sum_{j=R+2}^{N-R}\widetilde{\mu}_{NT,j}^{(k)}\leq(N-2R-1)\widetilde{\mu}_{NT,R+2}^{(k)}. \tag{52}$$

Then (52) implies that $B_1\leq V^{(k)}(R+1)\leq B_2$. By Lemma 3, $B_1=(N-3R-1)o_p(1)=o_p(N)$, and $B_2=(N-R-1)o_p(1)=o_p(N)$. Thus under Assumptions A – D, we have $V^{(k)}(R+1)=$

$o_p(N) \gg O_p(1)$. We now show the consistency of the GGR estimator. Using the inequalities

$$c/(1+c) < \ln(1+c) < c, \quad c \in (0, \infty), \tag{53}$$

we have that

$$\frac{\ln(1+\widetilde{\mu}^{(k)*}_{NT,j})}{\ln(1+\widetilde{\mu}^{(k)*}_{NT,j+1})} < \frac{\widetilde{\mu}^{(k)*}_{NT,j}}{\widetilde{\mu}^{(k)*}_{NT,j+1}/(1+\widetilde{\mu}^{(k)*}_{NT,j+1})} = \frac{\widetilde{\mu}^{(k)}_{NT,j}}{\widetilde{\mu}^{(k)}_{NT,j+1}} = O_p(1), \tag{54}$$

for $j = 1, 2, \ldots, R-1, R+1, \ldots, N-R-1$. Lemma 5, Lemma 6 and (54) imply that

$$\frac{V^{(k)}(R+1)}{V^{(k)}(R-1)} = \frac{V^{(k)}(R+1)}{\widetilde{\mu}^{(k)}_{NT,R} + \widetilde{\mu}^{(k)}_{NT,R+1} + V^{(k)}(R+1)} = O_p(1). \tag{55}$$

Using (53) and (55), we have that

$$\begin{aligned} \frac{\ln(1+\widetilde{\mu}^{(k)*}_{NT,R})}{\ln(1+\widetilde{\mu}^{(k)*}_{NT,R+1})} &> \frac{\widetilde{\mu}^{(k)*}_{NT,R+1}/(1+\widetilde{\mu}^{(k)*}_{NT,R+1})}{\widetilde{\mu}^{(k)*}_{NT,R}} \\ &= \frac{\widetilde{\mu}^{(k)}_{NT,R}}{\widetilde{\mu}^{(k)}_{NT,R+1}} \frac{V^{(k)}(R+1)}{V^{(k)}(R-1)} \to \infty. \end{aligned} \tag{56}$$

Then, the consistency of the GGR estimator follows from (54) and (56).

6.2. Jondeau et al. (2018)'s method

Jondeau et al. (2018) provide a simulation-based approach (JJR method later) to select the number of common factors that drive the variation in a sequence of cumulants. They use Monte Carlo simulations to derive the spectral densities of the correlation, co-skewness, and co-kurtosis tensors under the null hypothesis of independence, for large T and fixed N. Then the largest eigenvalue $\breve{\mu}_{NT,+}$ is used as a threshold to determine the number of common factors. That is $\widehat{R}_{JJR} = \#\{r \mid \tilde{\mu}_{NT,r} > \breve{\mu}_{NT,+}\}$ $(r = 1, 2, \ldots, N)$. We implement their approach as follows:

Step 1 For the dataset X with dimension N and T, we scale it to obtain normalized dataset Z. Then z_{it} is formalized with SGT distribution $SGT(z|\tau)$. Following Jondeau et al. (2018), the distribution parameter $\widehat{\tau}$ is estimated by fitting the average skewness and kurtosis of the data Z.

Step 2 We generate a series $\breve{z}_{it}$ with length NT, $\breve{z}_{it} \overset{i.i.d.}{\sim} SGT(z|\widehat{\tau})$. We, then conduct joint-MCA on the simulated dataset $\breve{Z}$ to obtain the maximum eigenvalue of the joint higher-order co-moment tensor, denoted as $\breve{\mu}_{NT,max}$.

Step 3 We repeat Step 2 l times and obtain the threshold as $\breve{\mu}_{NT,+} = l^{-1}\sum_{i=1}^{l}\breve{\mu}_{NT,i,max}$.

Step 4 We conduct joint-MCA on the original dataset Z to obtain the sample eigenvalue $\tilde{\mu}_{NT,r}(r = 1, 2, \ldots, N)$, the estimated number of non-Gaussian factors is $\widehat{R}_{JJR} = \#\{r \mid \tilde{\mu}_{NT,r} > \breve{\mu}_{NT,+}\}$.

We find that the result of the JJR method is not affected by increasing the value of l above 100. We, therefore, set $l = 100$ in our simulation studies and applications.

7. Computational aspects

It is well known that the dimension of the higher-order multi-cumulant tensor increases exponentially as the number of response series increases. For instance, the number of entries in the third-order multi-cumulant and fourth-order multi-cumulant of x_t are N^3 and N^4. When N is large, it is impossible to compute the higher-order multi-cumulant directly. HFA does not directly depend on the higher-order multi-cumulant but on its singular values. As shown in the main paper Lu et al. (2024), the singular value decomposition of $\widetilde{\mathbf{C}}_x^{(k)} \in \mathbb{R}^{N\times N^{k-1}}$ is equivalent to the teigenvalue decomposition of $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime} \in \mathbb{R}^{N\times N}$. Further, the computation of $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime}$ can be connected with the data matrix X but not $\widetilde{\mathbf{C}}_x^{(k)}$. Notice that for the k-th-order sample co-moment matrix $\widetilde{\mathbf{M}}_x^{(k)}$ of X, we have $\widetilde{\mathbf{M}}_x^{(k)}\widetilde{\mathbf{M}}_x^{(k)\prime} = T^{-2}X'[(X\ X')^{\circ(k-1)}]X$. When $k = 3$, it follows that

$$\widetilde{\mathbf{C}}_x^{(3)}\widetilde{\mathbf{C}}_x^{(3)\prime} = \widetilde{\mathbf{M}}_x^{(3)}\widetilde{\mathbf{M}}_x^{(3)\prime} = \frac{1}{T^2}X'((XX')\circ(XX'))X, \tag{57}$$

where $\circ$ denotes the Hadamard product. Since Equation (57) only contains the matrix computation, it is easy for computation and storage with `R` language. When $k = 4$, we denote $\widetilde{\Sigma}_x = X'X/T$ and use the definition of $\bar{\mathcal{H}}_x^{(4)}$ in Remark 4.2 of the main paper Lu et al. (2024). With simple algebra calculation, we have

$$\widetilde{\mathbf{C}}_x^{(4)}\widetilde{\mathbf{C}}_x^{(4)\prime} = \frac{1}{T^2}X'((XX')\circ(XX')\circ(XX'))X + \mathcal{N}_1 + \mathcal{N}_2 + \mathcal{N}_3, \tag{58}$$

where $\mathcal{N}_1 = -3(X'((b + b')\circ(XX'))X)/T^2$, $b = (a, a, \ldots, a) \in \mathbb{R}^{T\times T}, a = (a_1, a_2, \ldots, a_T)'$, $a_t = \sum_i\sum_j x_{it}x_{jt}\widetilde{\Sigma}_{x,ij}$ for $t = 1, 2, \ldots, T$; $\mathcal{N}_2 = 3\text{vec}(\widetilde{\Sigma}_x)'\text{vec}(\widetilde{\Sigma}_x)\widetilde{\Sigma}_x\widetilde{\Sigma}_x$; $\mathcal{N}_3 = 6\widetilde{\Sigma}_x\widetilde{\Sigma}_x\widetilde{\Sigma}_x\widetilde{\Sigma}_x\widetilde{\Sigma}_x$. The main computation cost is in $\mathcal{N}_1$, the vector a requires a loop for computation. Figure 2 compare the computation cost between SVD of $\widetilde{\mathbf{C}}_x^{(k)}$ and EVD of $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime}$ $(k = 3, 4)$. The computational cost of $\widetilde{\mathbf{C}}_x^{(k)}$ increases exponentially when N increases, in contrast, $\widetilde{\mathbf{C}}_x^{(k)}\widetilde{\mathbf{C}}_x^{(k)\prime}$ bears only very few cost even N is large.

∼ Insert Figure 2 Here ∼

The gain in accuracy of the HFA for estimating non-Gaussian factors comes at the cost of a higher (but feasible) computational cost. We document this in Figure 3. Since we fix $T = 500$ then increases N from 10 to 200. There is an increased computation time but the increase remains feasible in practice.

∼ Insert Figure 3 Here ∼

8. Additional simulations

8.1. Strong and weak non-Gaussian factors

The Monte Carlo simulations in the main paper assume all factors are either strong or weak. However, an empirical analysis shows that strong factors and weak factors are more likely to exist at the same time (Brown, 1989; Fama & French, 1993). Therefore, it is necessary to confirm the finite sample properties of the HFA estimators in a factor model with both strong and weak factors. We consider DGP1 in the main paper and all factors are non-Gaussian ($R = R_h = 2$). The distribution of the strong factor is skewed normal ($\eta_1 = 0.5, p_1 = 2, q_1 = \infty$; skewness $= 0.455$, ex-kurtosis $= 0.151$) and the weak factor is distributed as skewed-laplace ($\eta_2 = 0.5, p_2 = 1, q_2 = \infty$; skewness $= 1.244$, ex-kurtosis $= 4.920$). The corresponding factor loadings satisfies $\lambda_1 \sim \mathcal{N}(0, 2N^{-1/2})$ and $\lambda_2 \sim \mathcal{N}(0, N^{-1/2})$. The explanatory power of the first factor is twice that of the second factor. The serial correlation coefficient of factors $d = (0.5, 0.2)'$ and u_t draw from normal distribution and $\xi = 0.2$. The sample size is set as $(N = 100, T = 1000)$.

First, we evaluate the performance by determining the number of factors. We let $\sigma_j(G_N) = cj^{-1}$ for $j = 1, 2, \ldots, N$. Now we control the parameter c to adjust the relatively strength between factors and errors. To be specific, we set $c = 1, 2, 3, \ldots, 10$. We compute the eigenvalue ratio based on the covariance matrix and the third-order multi-cumulant of x_t to construct the ER and GER3 estimator. Figure 6(a1) and Figure 6(a2) report the eigenvalue ratio of ER and GER3 and the estimated number of factors, all results are averaged by 500 simulations. The spike of the eigenvalue ratio in Figure 6(a1) changes when c increases, however, the spike of the eigenvalue ratio in Figure 6(a2) remains the same. This indicates that the number of factors estimated by the ER estimator changes from two to one and the weak factor cannot be detected when the variance of the idiosyncratic errors increases. However, the GER3 estimator can still detect the weak factor. Figure 6(a3) shows the average number of factors estimated by ER, GR and GER3, we can also see that the ER and GR estimator can only detect the strong factor but GER3 can detect both strong and weak factors with c increases, which confirms the illustration in Section 3.2 of Lu et al. (2024).

Second, we need to evaluate the performance of the HFA estimators on the estimated strong and weak factors, respectively. We compute the Trace Ratio for the estimated strong and weak factors, respectively. We use DGP1 to control $\alpha \in [0, 1]$. Figure 6(b1) and Figure 6(b2) show the average Trace Ratio of the estimated strong and weak factors and the corresponding factor loadings, respectively. When α is close to one, namely Onatski (2012)'s weak factor model, the weak factor and corresponding factor loadings estimated by PCA become inefficient as shown before by Onatski (2012). In contrast, HFA show better performance on estimating the weak factors and corresponding factor loadings. In addition, Figure 6(b1) and (b2) show that PCA provides good performance for the strong factor.

$\sim$ Insert Figure 6 Here $\sim$

8.2. Non-Gaussian and Gaussian factors

When there are both Gaussian and non-Gaussian factors, we propose to use a two-step estimation of the number of factors, and an iterative approach for the estimation of the factors in Section 3. In this subsection, we verify the finite sample accuracy of this approach. We consider DGP1 ($R = 2$) in the main paper and one factor is non-Gaussian while another is Gaussian. u_t is normal distributed and $\xi = 0.2$. The serial correlation coefficient of factors $d = (0.5, 0.2)'$. We estimate the number of non-Gaussian factors by GER3 and estimate the number of Gaussian factors by ER and GR. The non-Gaussian and Gaussian factors and the corresponding factor loadings are estimated by the ALS algorithm with R_h and R_g being known.

As shown in Figure 7, the estimation procedure works well when both non-Gaussian and Gaussian factors exist. Figure 7(a) shows that GER3 can select the number of non-Gaussian factors R_h accurately as $\alpha \in [0, 1]$, the ER and GR estimators can select the number of Gaussian factors R_g when $\alpha < 0.6$, these are expected because, as shown in Section 3.2 of Lu et al. (2024), the ER or GR estimator cannot determine the accurate number of factors when the eigenvalue of G_N is large. The estimated non-Gaussian and Gaussian factors and the corresponding factor loadings, as shown in Figure 7(b), also confirm our theorems. The estimated Gaussian factor cannot hold the consistency when α is close to one, but it works in a strong factor model.

$\sim$ Insert Figure 7 Here $\sim$

8.3. Symmetric non-Gaussian factors

To study the sensitivity of the HFA approach for symmetric factors we consider the simple case in which, if all the non-Gaussian factors are symmetric and independent, then their third-order multi-cumulant would be a zero matrix (or $\phi_j^{(3)} = 0$ for $j = 1, 2, \ldots, R_h$), and further, if all the

non-Gaussian factors show a t distribution (fourth-order cumulants exist), then $0 < \phi_j^{(4)} < \infty$ for $j = 1, 2, \ldots, R_h$. The number of nonzero eigenvalues of $\mathbf{C}_f^{(k)}$ may be related with the order k.

We use simulation studies to investigate the case where the factors are symmetric. The distribution of the non-Gaussian factors now changes into Laplace distribution ($\eta_j = 0, p_j = 1, q_j = \infty$; skewness = 0, ex-kurtosis = 3). u_t is a normal distribution and $\xi = 0.2$. The serial correlation coefficient of factors is $d = (0.5, 0.2)'$. The number of factors remains 3 ($R = R_h = 3$). We first estimate the number of factors R using the GER and GGR estimators, then estimate the factors and factor loadings. For factor number selection, we consider the GER estimator based on the third- and fourth-order cumulant, and denote GER3 and GER4, respectively. For factors and the factor loadings estimation procedure, we consider the HFA estimators with $k = 3$ and $k = 4$ and denote as HFA3 and HFA4, respectively. When we estimate the factors and factor loadings, we assume the number of factors R is known. We set the maximum number of factors $R_{\max} = 10$.

Figure 8(a) and Figure 8(b) report the finite sample properties of GER3 and GER4 estimators. In sample size $(N, T) = (100, 1000)$ and $(N, T) = (100, 2000)$, we can observe that GER3 are inconsistent estimators of R. In contrast, GER4 estimators can choose the number of factors correctly even when α is close to one. When $(N, T) = (100, 1000)$, the GER4 estimator have poor finite sample properties when $\alpha = 1$, we need a larger sample size to support the accuracy as shown in Figure 8(b), this is expected because, as shown in Theorem 1, the statistics based on kurtosis have bad finite sample properties and need very large observations to obtain high power. On the other hand, see in Figure 8(c), HFA4 estimators remain consistent for all α, PCA and HFA3 estimators lose efficiency when α is close to one. It should be noted that we do not need a very large sample size for HFA4 to estimate factors and factor loadings consistently, therefore, if we have a priori information on the number of factors in a weak factor model, we can still use HFA4 to obtain a more efficient estimation than PCA.

∼ Insert Figure 8 Here ∼

9. Additional application results

This section gives the results of the normality test of the error terms in the main paper and the measurement of the decay rate of the spectrum of the error terms. We use the distribution-free normality test proposed by Bai & Ng (2005). Table 1 gives the proportion of rejecting the null hypothesis in the FRED-MD dataset. After we extracted 1 to 8 factors with HFA3 or HFA4, the majority of idiosyncratic errors cannot reject the normality test, which provides empirical evidence for the normality assumption of the main paper.

∼ Insert Table 1 Here ∼

Similarly, we measure the decay rate of the spectrum of G_N after extracting 1 to 8 factors with HFA3 or HFA4. We fit the eigenvalues by a polynomial decaying curve such that $\sigma_j(G_N) = C_0 j^{-\rho}$, where ρ is what we called "the decay rate of the spectrum". As shown in Table 2, the decay rate ρ is $0.618 \sim 0.812$ based on HFA3 approach and $0.753 \sim 0.926$ based on HFA3 approach. Even after extracting 8 factors, the decay rate ρ is still significantly larger than zero, which effectively supports the results in the main paper.

$\sim$ Insert Table 2 Here $\sim$

Finally, we test the normality of the error terms in the rolling-window analysis. As shown in Figure 9, the test result is robust. Only 6% of the idiosyncratic errors at most reject the normality test.

$\sim$ Insert Figure 9 Here $\sim$

10. Dampening of non-normality

In this section, we show that $e_t = G_N^{1/2} u_t$ being asymptotically normal according to the Lyapunov Central Limit Theorem (CLT) under mild assumptions on $G_N^{1/2}$. The Lyapunov Central Limit Theorem in Billingsley (2013) (Theorem 7.3) is as follows:

Theorem 7.3 of Billingsley (2013). *Suppose $\{X_1, \ldots, X_n\}$ $(n \to \infty)$ is a sequence of independent random variables, each with finite expected value μ_i and σ_i^2. Define*

$$s_n^2 \quad = \quad \sum_{i=1}^{n} \sigma_i^2 \, .$$

If for some $\delta > 0$, Lyapunov's condition

$$\lim_{n \to \infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{N} E\left[|X_i - \mu_i|^{2+\delta} \right] \quad = \quad 0$$

is satisfied, then

$$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \to_d \mathcal{N}(0,1) \, .$$

Now we use this theorem to show that when the number of non-zero elements in $G_N^{1/2}$ diverges to infinity as $N \to \infty$, we have the normality of $e_{1t} = \sum_{i=1}^{N} g_{1i} u_{it}$ under mild assumptions, where $(g_{11}, \ldots, g_{1N})$ is the first row of $G_N^{1/2}$. Define $\mathcal{D} = \{i : g_{1i} \neq 0\}$ and $\mathcal{D}$ indicates the index set of non-zero elements and we assume that $b_1 < \sqrt{|\mathcal{D}|}|g_{1i}| < b_2$ for $i \in \mathcal{D}$, where $|\mathcal{D}|$ is the number of

non-zero elements and b_1, b_2 are universal constants. Define $\mathcal{D}^c = \{i : g_{1i} = 0\}$ and $\mathcal{D}^c$ indicates the index set of zero elements. We assume $\mathcal{D} \cup \mathcal{D}^c = \{1, \ldots, N\}$. Let $X_{it} = g_{1i}u_{it}$, then we have $e_{1t} = \sum_{i=1}^{N} X_{it}$. Now we need to check whether $\{X_{1t}, \ldots, X_{Nt}\}$ satisfies the Lyapunov's condition. We also assume that $\{u_{it}\}_{t=1}^{T}$ is a stationary sequence (or simply consider *i.i.d.* u_{it}) for each i such that $E[u_{it}] = 0$, $E[u_{it}^2] = 1$ and $E[|u_{it}|^{2K}] < \infty$, and that $\{u_{it}\}_{t=1}^{T}$ are independent across i. We then have $\mu_i = E[X_{it}] = 0$ and $\sigma_i^2 = E[X_{it}^2] = g_{1i}^2$, $E[|X_{it}|^{2+\delta}] = |g_{1i}|^{2+\delta}E[|u_{it}|^{2+\delta}] \leq c_2|g_{1i}|^{2+\delta}$ by assuming $c_1 \leq E[|u_{it}|^{2+\delta}] \leq c_2$ for some universal positive constants c_1 and c_2 and any $\delta \leq 2K - 2$. It follows that

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{s_N^{2+\delta}} \sum_{i=1}^{N} E\big[|X_i - \mu_i|^{2+\delta}\big] &= \lim_{N\to\infty} \frac{1}{s_N^{2+\delta}} \sum_{i\in\mathcal{D}} E\big[|X_i - \mu_i|^{2+\delta}\big] \\
&\leq \lim_{N\to\infty} \frac{c_2 \sum_{i\in\mathcal{D}} |g_{1i}|^{2+\delta}}{\left(\sum_{i\in\mathcal{D}} g_{1i}^2\right)^{\frac{2+\delta}{2}}} \\
&\leq \lim_{N\to\infty} \frac{c_2|b_2|^{2+\delta}}{|b_1|^{2+\delta}} |\mathcal{D}|^{-\delta/2}.
\end{aligned}$$

If $|\mathcal{D}|$ (the number of non-zero elements) diverges to infinity as $N \to \infty$, the Lyapunov's condition is satisfied. On the other hand, the variance of e_{1t}, $s_N^2 = \sum_{i=1}^{N} g_{1i}^2 = \sum_{i\in\mathcal{D}} g_{1i}^2 \leq |b_2|^2 < \infty$. Overall, we can say e_{1t} is asymptotically normal distributed. This argument is analogous for e_{jt}, $j = 1, \ldots, N$.

11. Implementation in the `R` package `hofa`

The main functions of HFA in the package `hofa` are `M3.select`, `M4.select`, `M3.als` and `M4.als`. The functions `M3.select` and `M4.select` provide Generalized Eigenvalue Ratio (GER) test for determining the number of non-Gaussian and Gaussian factors, and the function `M3.als` and `M4.als` implement the Alternating Least Square (ALS) algorithm to estimate the factors and factor loadings in Higher-order multi-cumulant factor analysis.

Installing and loading the package from GitHub are achieved by

```
> devtools::install_github("GuanglinHuang/hofa")
> library("hofa")
```

The function `M3.select` (also for `M4.select`) takes as arguments the $T \times N$ dataset of T observations of the N-variate variable $\boldsymbol{X}$. Optional arguments include, a logical parameter `scale`, whether the dataset need to be normalized (default of `FALSE`), the maximum number of factors `rmax` (default of 8), and the method be used: `"GER3"`, Generalized Eigenvalue Ratio test based

on third-order cumulant; "GGR3", Generalized Growth Ratio test based on third-order cumulant; "JJR", Jondeau et al. (2018)'s threshold approach. The documentation of the M3.sel function can be loaded by

```
> ?M3.sel
```

We use the edhec dataset, which contains monthly returns returns on the Convertible Arbitrage and CTA global EDHEC hedge fund style indices over the period January 1997 until November 2019, to illustrate this function.

```
> data(edhec)
> data = edhec[, 1:13]*100
> fn_ger3 <- M3.select(data, method = "GER3")
> fn_ger4 <- M4.select(data, method = "GER4")
> fn_er <- M2.select(data, method = "ER")
> names(fn_ger3)
[1] "R"           "Rh"          "Rg"          "eigenvalues"
> names(fn_ger4)
[1] "R"           "Rh"          "Rg"          "eigenvalues"
> names(fn_er)
[1] "R"           "eigenvalues"
```

The approach "GER3" and "GER4" return the estimated number of non-Gaussian (R_h), Gaussian factors (R_g), all factors ($R = R_h + R_g$) and the eigenvalues of the higher-order multi-cumulant. The approach "ER" reports the estimated total number of factors (R) and the eigenvalues of the covariance matrix. The results of those approaches are as follows

```
> fn_ger3$R
[1] 3
> fn_ger3$Rh
[1] 1
> fn_ger3$Rg
[1] 2
> fn_ger4$R
[1] 4
> fn_ger4$Rh
[1] 2
```

```
> fn_ger4$Rg
[1] 2
> fn_er$R
[1] 1
```

For example, the estimated number of non-Gaussian factors $\widehat{R}_h$ is one by GER3 and two by GER4, the estimated number of Gaussian factors $\widehat{R}_g$ is two by both ER and GR criterions. The estimated number of factors $\widehat{R}$ are 3,4 and 1 by GER3, GER4 and ER, respectively.

The function `M3.als` (also for `M4.als`) estimates the factor loadings and factors by giving `Rh`, `Rg` and the $T \times N$ dataset of $\boldsymbol{X}$. As explained and demonstrated in the main paper, given consistent estimates of the number of non-Gaussian factors R_h and Gaussian factors R_g, we can estimate the factors and factor loadings consistently. Therefore, without a priori information of the factor number, the results of the `M3.select` or `M4.select` function can be used. Optional arguments of `M3.als` include (i) a logical parameter `scale`, whether the dataset need to be normalized (default of `FALSE`) and (ii) the the iteration error `eps` (default to 10^{-8}).

Continuing with the example of the `edhec` dataset. By setting `rh = 2` and `rg = 2`, we can estimate the factors and the factor loadings as follows

```
> est_hfa3 = M3.als(data, rh = 2, rg = 2)
> est_hfa4 = M4.als(data, rh = 2, rg = 2)
> names(est_hfa3)
[1] "f"  "u"  "e"  "ev"
> round(head(est_hfa3$f),2)
      [,1]  [,2]  [,3]  [,4]
[1,]  1.88 -0.68 -1.77 -1.20
[2,] -0.01 -1.06 -0.59 -1.53
[3,] -1.81 -1.10  0.83 -0.73
[4,]  0.37  0.07  0.33  0.09
[5,]  2.44  0.32 -0.48  0.74
[6,]  1.55 -0.73 -0.48 -0.81
> round(head(est_hfa4$f),2)
      [,1]  [,2]  [,3]  [,4]
[1,]  1.94 -0.61 -1.79 -1.10
[2,] -0.01 -0.99 -0.73 -1.47
[3,] -1.88 -1.00  0.66 -0.87
[4,]  0.38  0.02  0.38  0.08
```

```
[5,]  2.45  0.28 -0.36  0.72
[6,]  1.55 -0.67 -0.52 -0.85
```

The first `rh` columns are the estimated non-Gaussian factors and the remaining `rg` columns are the estimated Gaussian factors. Ignore the signs of the factors, the results are different between `HFA3` and `HFA4`. Moreover, the factors are different when Principal Component Analysis (PCA) is used

```
> est_pca = M2.pca(data,r = 4, method = "PCA")
> round(head(est_pca$f),2)
      [,1]  [,2]  [,3]  [,4]
[1,]  2.08 -1.79 -0.57  0.92
[2,] -0.10 -1.74  0.13  0.87
[3,] -2.09 -0.83  0.81  0.01
[4,]  0.32  0.27  0.36  0.11
[5,]  2.57  0.37 -0.03 -0.40
[6,]  1.49 -1.12  0.40  0.43
```

References

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

Anderson, T., & Gupta, S. D. (1963). Some inequalities on characteristic roots of matrices. *Biometrika*, *50*, 522–524.

Ando, T., & Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, *31*, 163–191.

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., & Ng, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business and Economic Statistics*, *23*, 49–60.

Bai, J., & Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, *74*, 1133–1150.

Bai, J., & Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, *176*, 18–29.

Bai, J., & Ng, S. (2023). Approximate factor models with weaker loadings. *Journal of Econometrics*, .

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Boudt, K., Cornilly, D., & Verdonck, T. (2020). Nearest comoment estimation with unobserved factors. *Journal of Econometrics*, *217*, 381 – 397.

Brown, S. J. (1989). The number of factors in security returns. *The Journal of Finance*, *44*, 1247–1262.

Cardoso, J. F., & Souloumiac, A. (1993). Blind beamforming for non-gaussian signals. *IEE proceedings F (Radar and Signal Processing)*, *140*, 362–370.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, *21*, 1324–1342.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*, 3–56.

Freyaldenhoven, S. (2022). Factor models with local factorsdetermining the number of relevant factors. *Journal of Econometrics*, *229*, 80–102.

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*, 1483–1492.

Jondeau, E., Jurczenko, E., & Rockinger, M. (2018). Moment component analysis: An illustration with international stock markets. *Journal of Business and Economic Statistics*, *36*, 576–598.

Lu, W., & Huang, G. (2022). Estimating the higher-order co-moment with non-gaussian components and its application in portfolio selection. *Statistics*, *56*, 537–564.

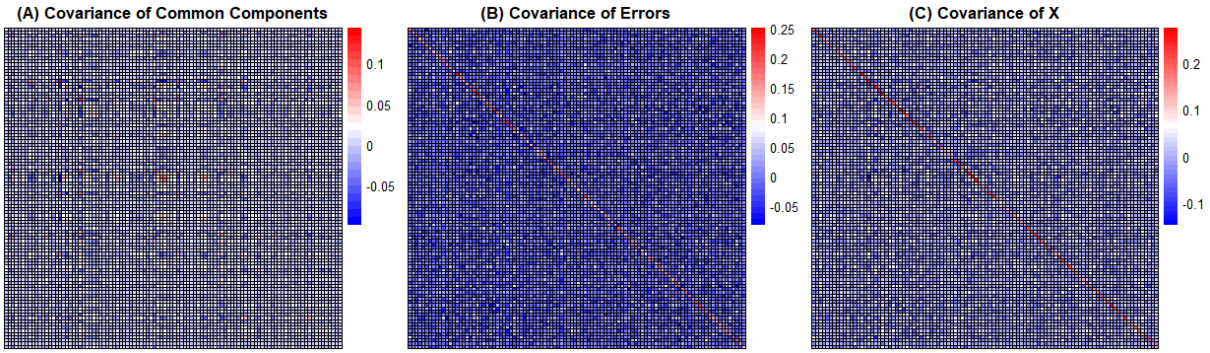
Lu, W., Huang, G., & Boudt, K. (2024). Estimation of non-gaussian factors using higher-order multi-cumulants in weak factor models. *Working paper*, *Available at SSRN*, https://ssrn.com/abstract=3864960.

McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business and Economic Statistics*, *34*, 574–589.

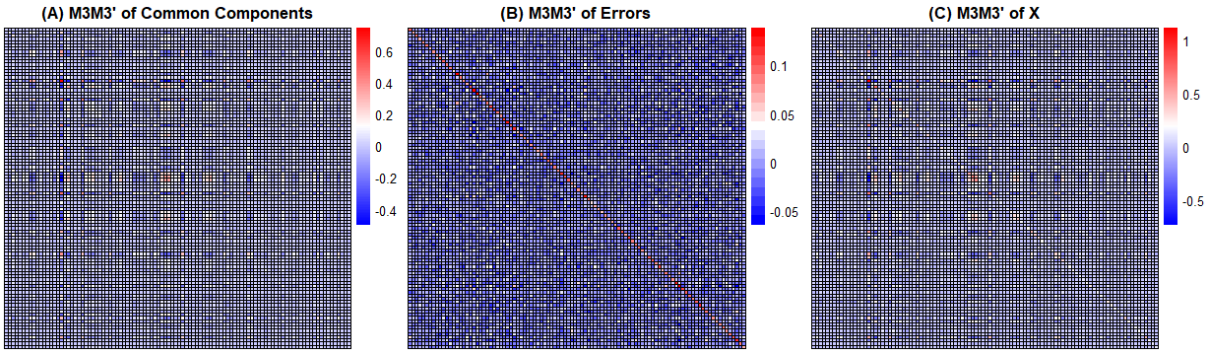
Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, *92*, 1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, *168*, 244–258.

Saulis, L., & Statulevicius, V. (1991). *Limit theorems for large deviations* volume 73. Springer Science & Business Media.



(a) The heat map structure based on covariance matrix



(b) The heat map structure based on third-order multi-cumulant

Figure 1: The heat map of the covariance matrix and third-order multi-cumulant of c_{it}, e_{it} and x_{it}

Note: This figure reports the elements in covariance matrix and third-order multi-cumulant of c_{it}, e_{it} and x_{it}. The DGP follows the DGP1 in the main paper and the sample size is $(N, T) = (300, 500)$ and $\alpha = 1$.

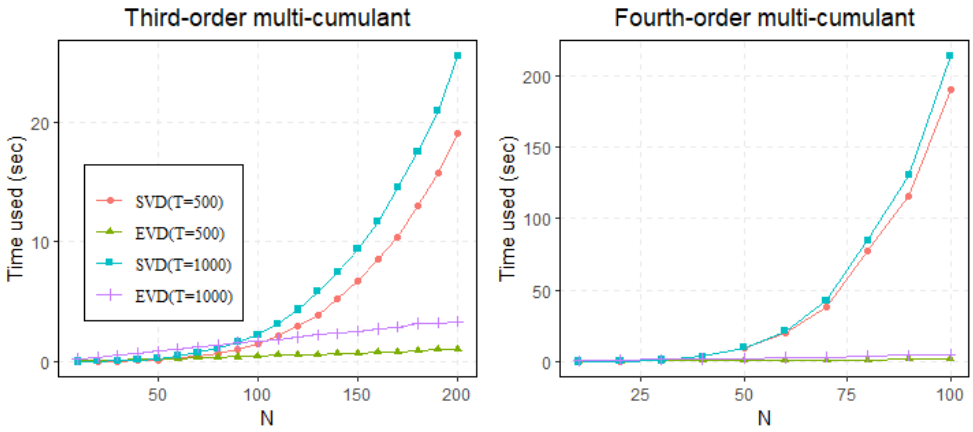


Figure 2: Computational cost of singular values of higher-order multi-cumulant

Note: This Figure reports the computation cost of singular value decomposition of higher-order multi-cumulant. SVD denotes compute the singular value decomposition on $\tilde{\mathbf{C}}_x^{(3)}(\tilde{\mathbf{C}}_x^{(4)})$ directly, EVD denotes compute eigenvalue decomposition on $\tilde{\mathbf{C}}_x^{(3)}\tilde{\mathbf{C}}_x^{(3)\prime}(\tilde{\mathbf{C}}_x^{(4)}\tilde{\mathbf{C}}_x^{(4)\prime})$. We have 10 replications.

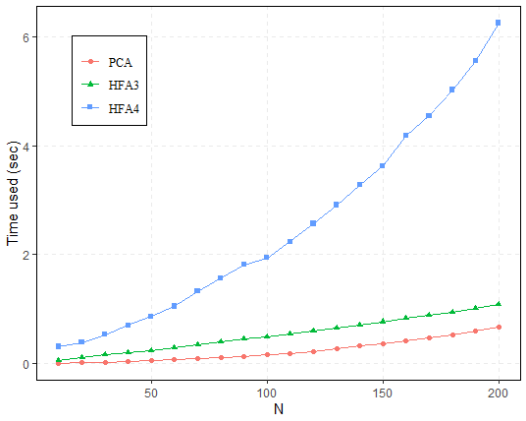
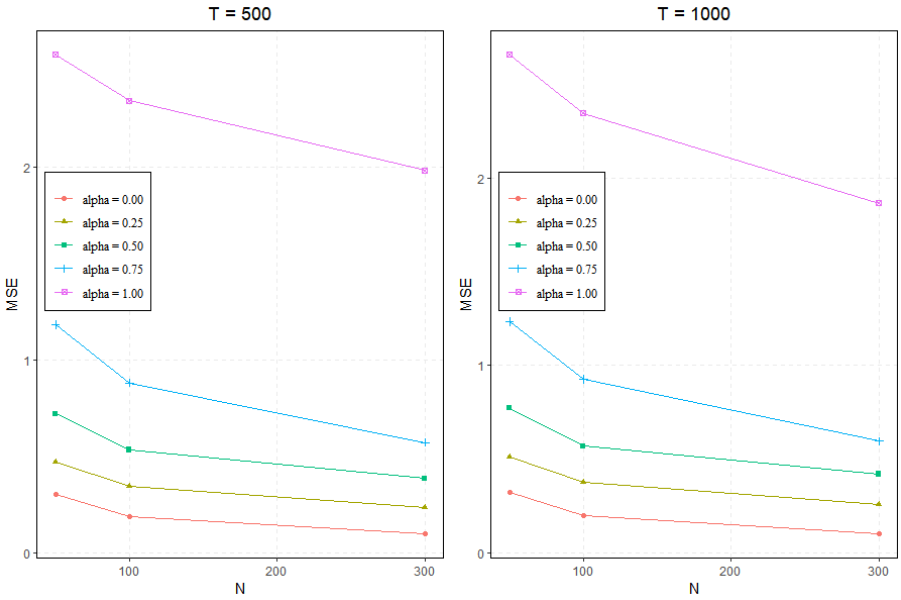
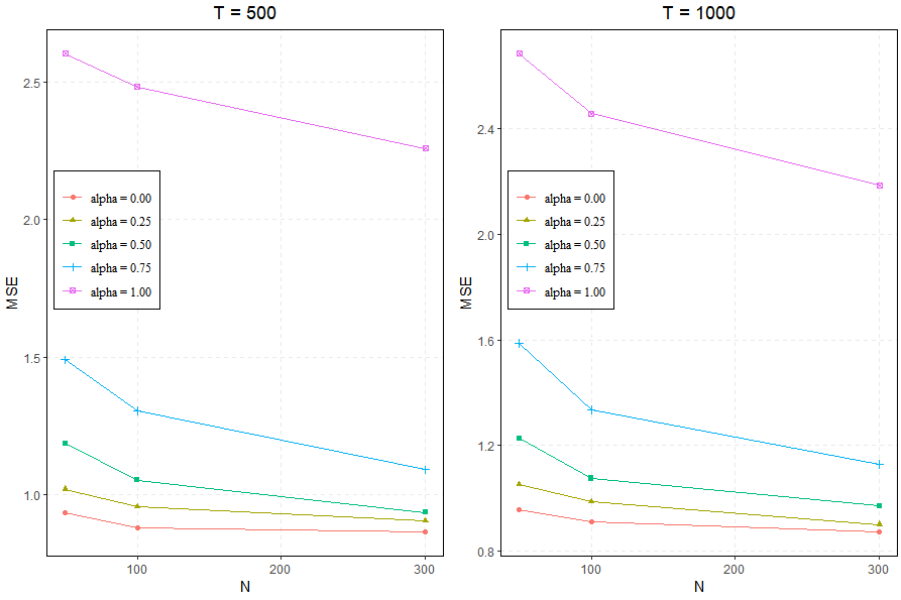


Figure 3: Computational cost of HFA3, HFA4 and PCA

Note: This Figure reports the computation cost of singular value decomposition of HFA3, HFA4 and PCA. HFA3 denotes compute eigenvalue decomposition on $\tilde{\mathbf{C}}_x^{(3)}\tilde{\mathbf{C}}_x^{(3)\prime}$, HFA4 denotes compute eigenvalue decomposition on $\tilde{\mathbf{C}}_x^{(4)}\tilde{\mathbf{C}}_x^{(4)\prime}$, and PCA denotes compute eigenvalue decomposition on $\tilde{\mathbf{C}}_x^{(2)}$. We have 10 replications.



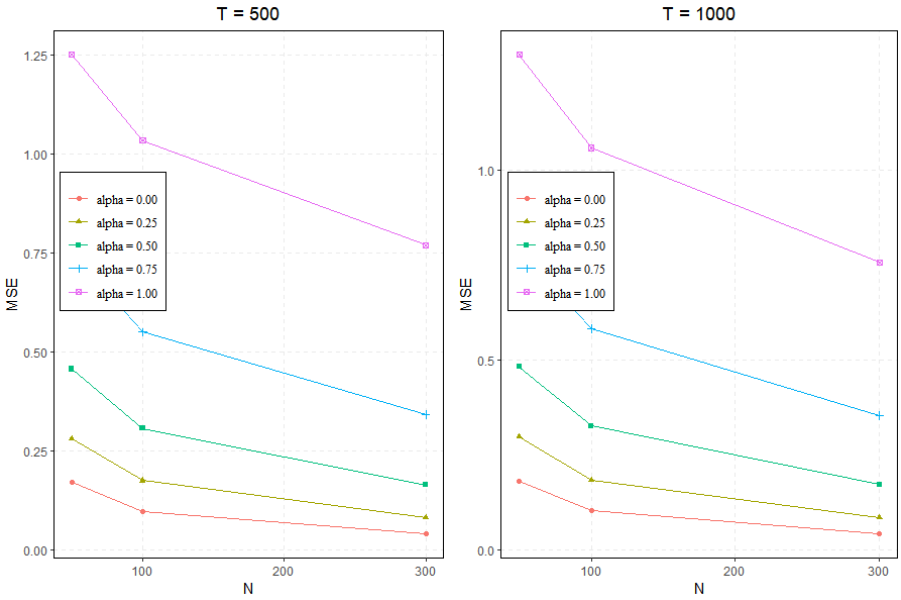
(a) The finite sample properties of estimated differenced factor $\widehat{\Delta f}_t$



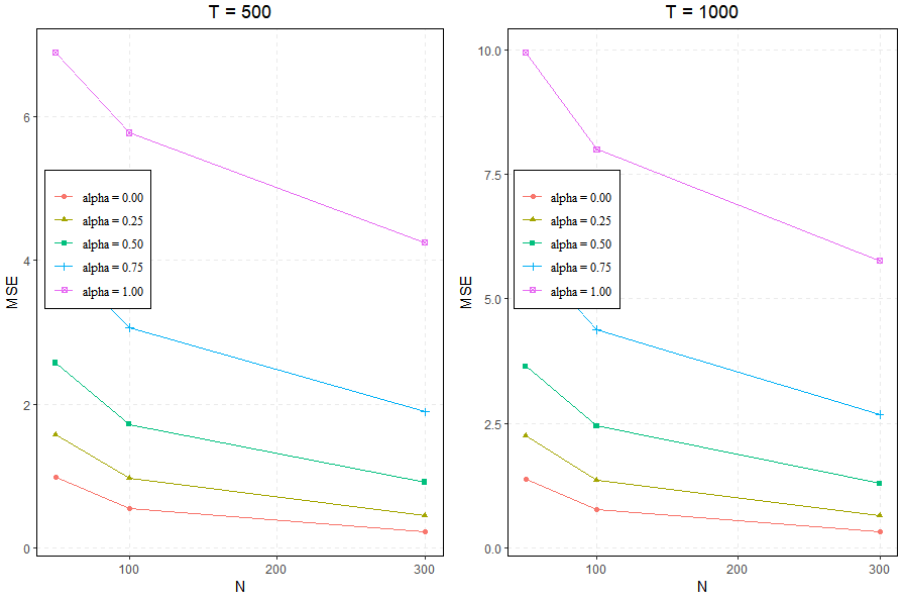
(b) The finite sample properties of estimated factor $\widehat{f}_t$

Figure 4: Accuracy of HFA factors with a time trend

Note: This Figure reports the maximum square error of the HFA estimation with the non-stationary data which has a time trend, see in (40). Figure (a) reports the maximum square error between $\widehat{\Delta f}_t$ and Δf_t, where $\widehat{\Delta f}_t$ is estimated by HFA using the first order differenced data. Figure (b) reports the maximum square error between $\widehat{f}_t$ and f_t. The estimated factor $\widehat{f}_t$ is recovered by the cumulative sum of the estimated differenced factor $\widehat{\Delta f}_t$. We have 2000 replications.



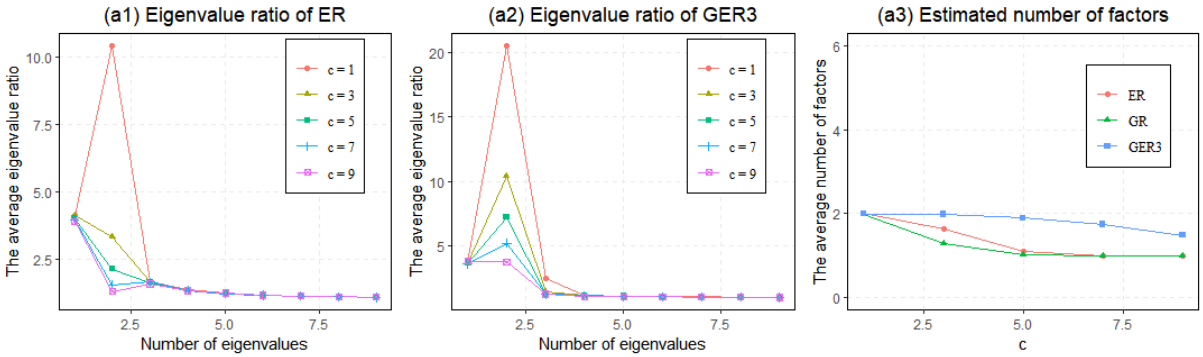
(a) The finite sample properties of estimated differenced factor $\widehat{\Delta f}_t$



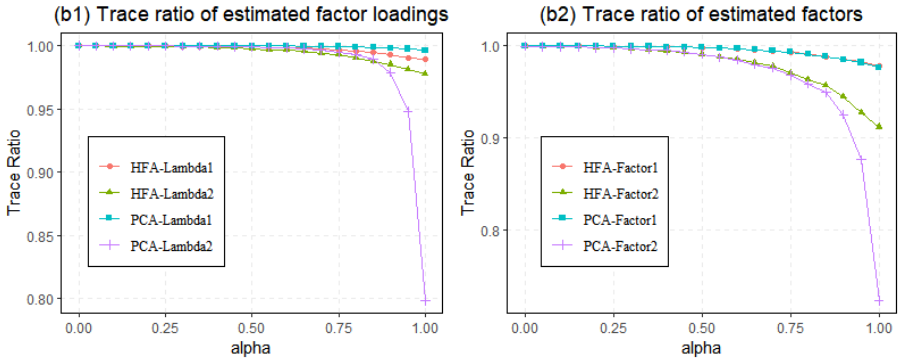
(b) The finite sample properties of estimated factor $\widehat{f}_t$

Figure 5: Accuracy of HFA factors with I(1) factor and I(1) idiosyncratic errors

Note: This Figure reports the maximum square error of the HFA estimation with the non-stationary data which both f_t and e_{it} are I(1) process, see in (44). Figure (a) reports the maximum square error between $\widehat{\Delta f}_t$ and Δf_t, where $\widehat{\Delta f}_t$ is estimated by HFA using the first order differenced data. Figure (b) reports the maximum square error between $\widehat{f}_t$ and f_t. The estimated factor $\widehat{f}_t$ is recovered by the cumulative sum of the estimated differenced factor $\widehat{\Delta f}_t$. We have 2000 replications.



(a) The finite sample properties of estimated factor number



(b) The finite sample properties of estimated factors and loadings

Figure 6: Accuracy of HFA and PCA when both strong and weak factors existing

Note: This figure reports the finite sample properties of HFA and PCA in a two-factor model with one strong factor and one weak factor. Both of two factors are non-Gaussian. We follow the DGP1 in the main paper and the sample size is $(N, T) = (100, 1000)$. Figure (a1) and (a2) show the average eigenvalue ratio of ER (GR) and GER3 estimators based on 500 simulations, Figure (a3) shows the average number of factors estimated by ER (GR) and GER3 estimators, where c controls the eigenvalues of the idiosyncratic errors. Figure (b1) and Figure (b2) show the Trace Ratio of the estimated factors loadings and factors, respectively.

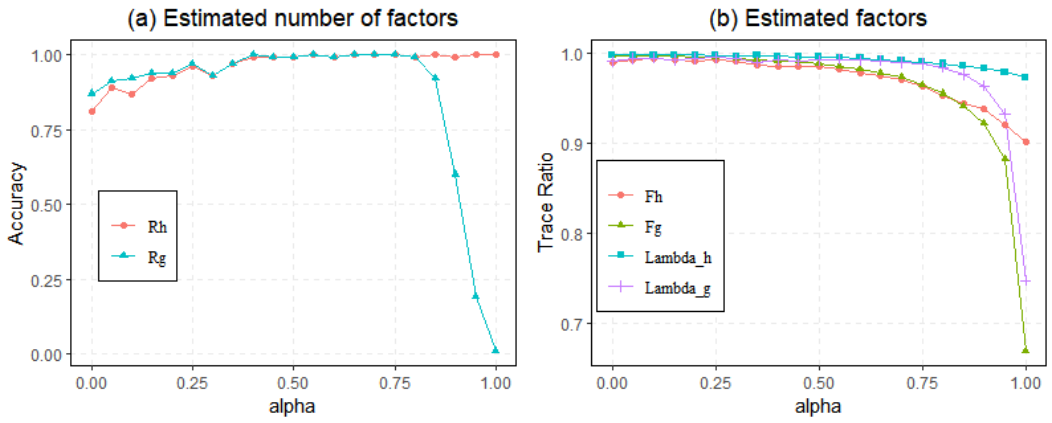


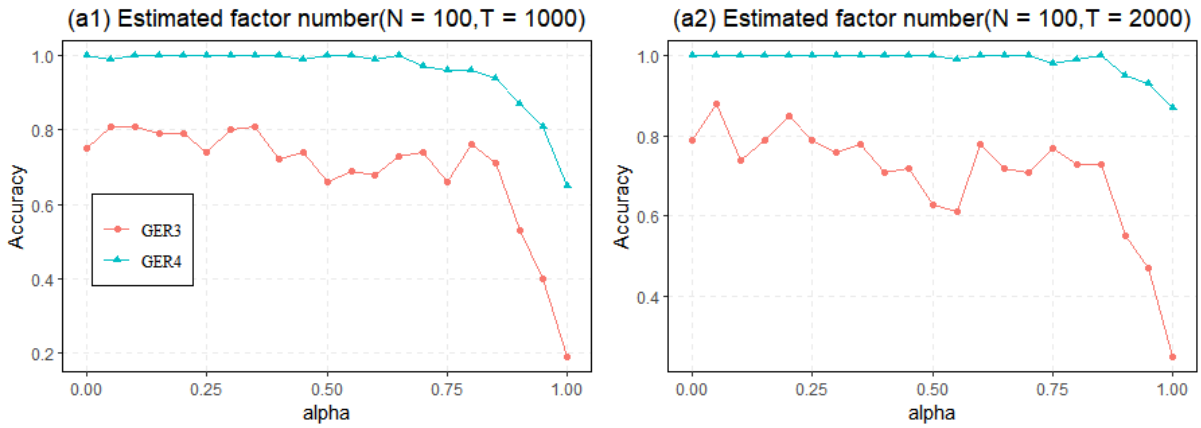
Figure 7: Accuracy of HFA estimators when both non-Gaussian and Gaussian factors existing

Note: This figure reports the finite sample properties of HFA estimators with both non-Gaussian and Gaussian factors existing. We follow the DGP1 in the main paper with one non-Gaussian factor and one Gaussian factor. Figure (a) show the proportion of selecting the number of non-Gaussian (R_h) and Gaussian (R_g) factors correctly by GER3 (GGR3) and ER (GR), respectively. Figure (b) shows the average Trace Ratio of non-Gaussian (F_h) and Gaussian (F_g) factors and the corresponding factor loadings (Λ_h and Λ_g) estimated by ALS algorithm. The sample size is $(N, T) = (300, 500)$ and we have 500 replications.

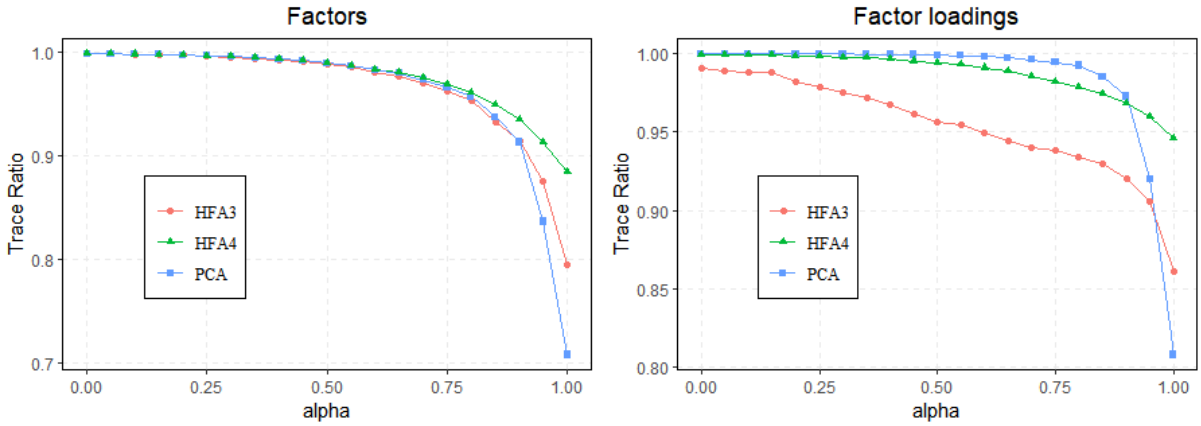
Table 1: The normality test of the idiosyncratic errors in FRED-MD dataset

	family-wise error rate = 10%			family-wise error rate = 5%			family-wise error rate = 1%		
	Normality	Asymmetric	Fat-tailed	Normality	Asymmetric	Fat-tailed	Normality	Asymmetric	Fat-tailed
Panel A: HFA3 approach									
$r = 1$	1.61	0.00	2.42	0.81	0.00	1.61	0.00	0.00	0.00
$r = 2$	3.23	0.00	1.61	2.42	0.00	0.81	0.00	0.00	0.00
$r = 3$	2.42	0.00	2.42	1.61	0.00	0.81	0.00	0.00	0.00
$r = 4$	3.23	0.00	2.42	2.42	0.00	1.61	1.61	0.00	0.81
$r = 5$	3.23	0.00	2.42	2.42	0.00	1.61	1.61	0.00	0.81
$r = 6$	3.23	0.00	2.42	2.42	0.00	1.61	0.81	0.00	1.61
$r = 7$	3.23	0.00	3.23	2.42	0.00	1.61	1.61	0.00	1.61
$r = 8$	6.45	0.00	3.23	4.84	0.00	2.42	0.81	0.00	2.42
Panel B: HFA4 approach									
$r = 1$	1.61	0.00	3.23	0.81	0.00	1.61	0.00	0.00	0.00
$r = 2$	2.42	0.00	1.61	0.81	0.00	0.81	0.00	0.00	0.00
$r = 3$	3.23	0.00	2.42	1.61	0.00	1.61	0.00	0.00	0.00
$r = 4$	1.61	0.00	2.42	1.61	0.00	1.61	0.00	0.00	0.00
$r = 5$	4.03	0.00	2.42	3.23	0.00	1.61	1.61	0.00	1.61
$r = 6$	3.23	0.00	2.42	3.23	0.00	1.61	1.61	0.00	1.61
$r = 7$	4.03	0.00	3.23	4.03	0.00	2.42	2.42	0.00	2.42
$r = 8$	4.03	0.81	3.23	3.23	0.00	2.42	2.42	0.00	2.42

Note: This table reports the normality test of idiosyncratic errors of the FRED-MD dataset after extracting several factors by HFA3 or HFA4 approach, respectively. r indicates the number of HFA components have been remove before the test. The numbers in the table represent the proportion(%) of rejecting the null hypothesis to the total ($N = 124$). "Asymmetric" and "Fat-tailed" represent the test of the third-order moment and the fourth-order moment, respectively.



(a) Accuracy of estimated factor number



(b) Accuracy of estimated factors and loadings (N = 100, T = 1000)

Figure 8: Accuracy of HFA estimators with symmetric factors

Note: This figure reports the finite sample properties of HFA estimators with symmetric non-Gaussian factors. Figure (a1) and (a2) show the proportion of selecting the number of factors correctly by GER3 and GER4, Figure (b) shows the average Trace Ratio of factors and factor loadings estimated by HFA3, HFA4 and PCA. We have 500 replications.

Table 2: Measurement of the decay rate of the spectrum of the error terms in the FRED-MD dataset

	HFA3	HFA4
$r = 1$	0.812	0.926
$r = 2$	0.738	0.924
$r = 3$	0.674	0.910
$r = 4$	0.664	0.829
$r = 5$	0.634	0.826
$r = 6$	0.638	0.826
$r = 7$	0.634	0.826
$r = 8$	0.618	0.753

Note: This table reports the decay rate ρ in the FRED-MD dataset after extracting several factors by HFA3 or HFA4 approach, respectively. r indicates the number of HFA components have been remove. ρ is fitted by the polynomial decaying function $\sigma_j(G_N) = C_0 j^{-\rho}$.

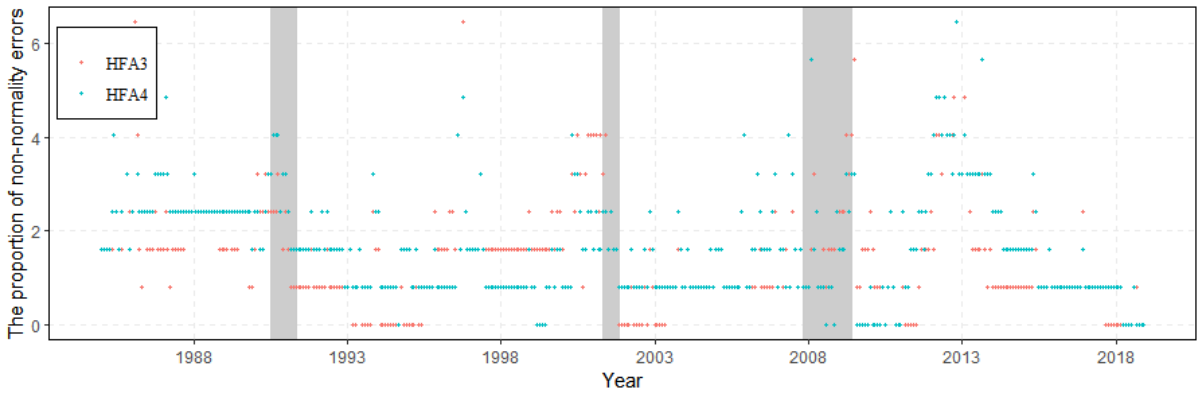


Figure 9: Rolling-window normality test of the idiosyncratic errors

Note: This figure reports the normality test of idiosyncratic errors of the FRED-MD dataset after extracting four factors by HFA3 or HFA4 approach, respectively. The points in the figure represent the proportion(%) of rejecting the null hypothesis to the total ($N = 124$). Shaded regions indicate the three largest drawdown periods of the S&P 500 during the out-of-sample period.