Spurious Factor Analysis

Alexei Onatski* and Chen Wang†
*Faculty of Economics, University of Cambridge.
†Department of Statistics and Actuarial Science,
University of Hong Kong.

June 17, 2020

Abstract

This paper draws parallels between the Principal Components Analysis of factorless high-dimensional nonstationary data and the classical spurious regression. We show that a few of the principal components of such data absorb nearly all the data variation. The corresponding scree plot suggests that the data contain a few factors, which is collaborated by the standard panel information criteria. Furthermore, the Dickey-Fuller tests of the unit root hypothesis applied to the estimated "idiosyncratic terms" often reject, creating an impression that a few factors are responsible for most of the non-stationarity in the data. We warn empirical researchers of these peculiar effects and suggest to always compare the analysis in levels with that in differences.

KEY WORDS: Spurious regression, principal components, factor models, Karhunen-Loève expansion.

1 Introduction

Researchers applying factor analysis to nonstationary macroeconomic panels face a choice: keep the data in levels or first-difference them. If all the nonstationarity is due to factors, no differencing is necessary. A simple principal components estimator of the factors is consistent and more efficient than that based on the differenced data (e.g. Bai, 2004). Otherwise, the standard advice is to extract the factors from the

first-differenced data, and then, accumulate them to obtain estimates of the factors in levels (e.g. Bai and Ng, 2004).

Both strategies are used in practice. For example, Moon and Perron (2007), Eickmeier (2009), Wang and Wu (2015), von Borstel et al. (2016), and Barigozzi et al. (2018) fit factor models to non-stationary data after first-differencing them. Stock and Watson (2016) not only first-difference most of the series entering their dynamic factor model of the US economy, but also locally demean the variables to minimize problems associated with low-frequency variability. On the other hand, Bai (2004), Corielli and Marcellino (2006), Ghate and Wright (2012), West and Wong (2014), and Engel et al. (2015) estimate factor models on non-stationary data in levels.

Factor estimation in levels relies on the assumption of stationary errors. Banerjee et al. (2017, section 4.1) give "several reasons for making the hypothesis of I(0) idiosyncratic errors" in macroeconomic applications. One of their reasons is a very high rejection rate of the hypothesis of a unit root in the estimated idiosyncratic components of the 114 nonstationary monthly US macroeconomic series for the 1959-2014 period (see their Footnote 5).

This paper is intended as a warning to the empirical researchers tempted by arguments advocating factor estimation in levels. We show theoretically that a few principal components of a *factorless* nonstationary panel must "explain" an extremely high portion of the data variation. Moreover, the Dickey-Fuller tests on the estimated idiosyncratic terms are strongly oversized, supporting the stationarity hypothesis where, in fact, the null of nonstationarity is true.

We are not the first to point out the high explanatory power of a few of the principal components of factorless persistent data. Uhlig (2009), discussing Boivin et al. (2009), generates artificial cross-sectionally *independent* AR(1) data with the autoregressive coefficients matching the first-order autocorrelations of the 243 macroeconomic series used in Boivin et al. (2009). Then he plots the fraction of variation explained against the number of factors for both actual and artificial data (see Figure 1), and notes that the two plots "look surprisingly and uncomfortably alike". In particular, five estimated factors explain about 75% of the actual data variation, but at the same time, five estimated factors, that must be spurious by construction, "explain" about 60% of the simulated data variation.

Uhlig (2009) attributes the high explanatory power of the spurious factors to the fact that the simulated data are considerably autocorrelated. Many of the simulated

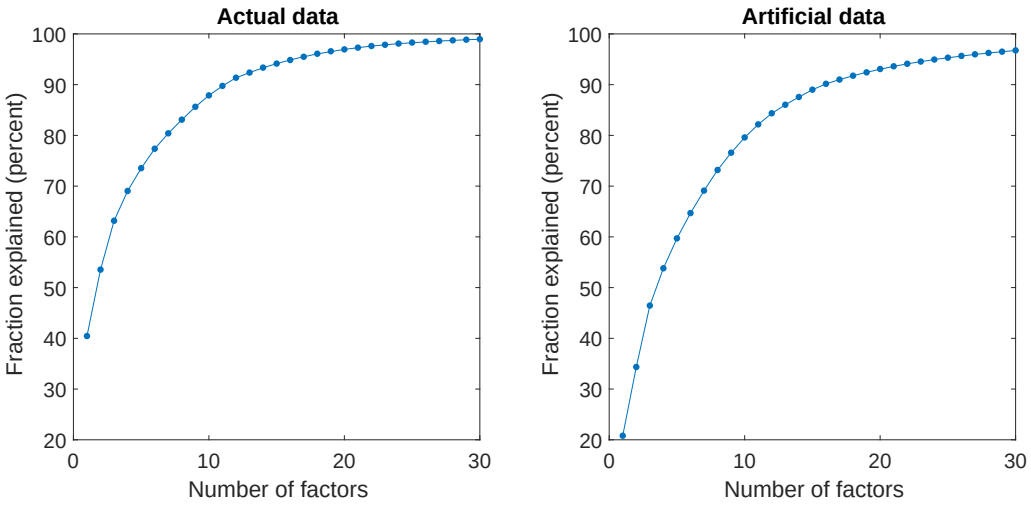

Figure 1: Factor contribution to the overall variance. Left panel: actual Boivin et al.'s (2009) data. Right panel: factorless simulated data with similar autocorrelation properties.

series' first-order autocorrelation coefficients are close to unity. In a finite sample (in his setting, 83 observations), the series may appear to be correlated, which will be picked up by the principal components. Although this explanation is intuitive, Uhlig admits that it is "perhaps tricky to formalize".

In this paper, we do such a formalization at different levels of generality. In our basic setting, the data are generated by a high-dimensional integrated system with an increasing number of stochastic trends, none of which is dominating the rest asymptotically. An extreme example would be a panel of cross-sectionally *independent* difference-stationary processes. The setting also covers more empirically relevant situations with any types of cross-sectional dependence except those caused by the presence of a fixed number of genuine strong nonstationary factors in the data.

We prove that in such a setting the fraction of the data variation explained by the first principal component converges in probability to $6/\pi^2 \approx 0.61$ even when the data do not contain any common factors. The first three principal components together asymptotically explain $100\% \sum_{j=1}^{3} 6/(j\pi)^2 \approx 83\%$ of the variation in the *factorless* nonstationary data. The corresponding "factor estimates" converge to deterministic cosine waves that resemble linear, quadratic, and cubic time trends.

The flavour of these results is preserved in a more general setting of a local level

model, where the data are represented by a weighted sum of I(1) and I(0) processes with the weights on the former possibly decaying to zero as the sample size increases. Furthermore, our conclusions do not change qualitatively when data are not only demeaned but also standardized before the principal components analysis (PCA) is applied.

We show from a theoretical standpoint that, in our basic setting, the standard panel information criteria (e.g. Bai, 2004) are very sensitive to the choice of the *a priori* maximum number of factors. For empirically relevant choices and data sizes, the criteria will often detect two or three "factors". We provide Monte Carlo evidence supporting this claim.

The peculiar results of the PCA of factorless nonstationary data are relatively easy to explain in the extreme case where the data are given by cross-sectionally i.i.d. random walks. In such a case, the sample covariance matrix used by the PCA to extract "factors" can be interpreted as a discrete time approximation for the covariance operator of a demeaned Wiener process. As the data dimensions grow, the PCA estimates of the "factors" converge to the eigenfunctions of the covariance operator, which happen to be the cosine waves. The explanatory power of the estimated "j-th factor" converges to the j-th largest eigenvalue of the covariance operator, which equals $6/\left(j\pi\right)^2$.

A somewhat different explanation relates to the Karhunen-Loève expansion of the demeaned Wiener process (e.g. Shorack and Wellner, 1986). The expansion represents the process in the form of an infinite sum of trigonometric functions with uncorrelated random coefficients whose variances are quickly decaying. Since difference-stationary series can be approximated by Wiener processes, much of the variation in a nonstationary panel can be captured by a few of the trigonometric functions corresponding to the first terms in the Karhunen-Loève expansion.

Phillips (1998) points out that the "prototypical spurious regressions, in which unit root nonstationary time series are regressed on deterministic functions," reproduces the underlying Karhunen-Loève representation of the Wiener process. In the similar spirit, the spurious factor analysis, i.e. the principal components analysis of factorless difference-stationary data, picks up the common Karhunen-Loève structure of the

cross-sectional units.[1]

This intuition immediately suggests that the Dickey-Fuller tests of the hypothesis of a unit root in the estimated spurious idiosyncratic terms must be oversized. Indeed, when researchers apply the test to an estimated idiosyncratic term, they ignore the fact that the estimate is, essentially, the residual from a regression of a nonstationary series on a few slowly varying trigonometric functions.

These functions are similar to the deterministic polynomial trends. Hence, the intercept-only Dickey-Fuller statistic computed on the basis of estimated idiosyncratic terms asymptotically behaves similarly to the intercept-only Dickey-Fuller statistic for the regression that includes several deterministic polynomial time trends. This leads to a substantial size distortion and a potentially confused conclusion that the factors soak up all or most of the nonstationarity in the data.

All in all, the results of the principal components analysis of the levels of nonstationary data may be very misleading. We recommend to always compare the first differences of factors estimated from the levels with factors estimated from the first-differenced data. A mismatch indicates a spurious factor analysis in levels.

The remainder of the paper is structured as follows. In Section 2, we formally introduce our setting and present our main results. Section 3 discusses various extensions to the basic setting. Section 4 studies the workings of the information criteria for the determination of the number of factors in the context of spurious factor analysis. Section 5 discusses ways to detect spurious results. Section 6 concludes. Technical discussions of the assumptions, Monte Carlo analysis, and the proof of our main result, Theorem 1, are reported in the Appendix. Proofs of the other results are given in the Supplementary Material (SM).

2 Basic setup and main results

Consider an N-dimensional integrated system

$$X_t = X_{t-1} + \Psi(L)\,\varepsilon_t, \tag{1}$$

[1]The notion of spurious factors considered in this paper is not directly related to the spurious factors in asset returns that received much recent research attention (see Bryzgalova (2018) and references therein).

where ε_t is N_ε-dimensional, and matrix $\Psi(1)$ may be of deficient rank so that cointegration is allowed. Suppose that data are summarized by the $N \times T$ matrix $X = [X_1, ..., X_T]$. Our goal is to study the workings of the PCA of these data as both N and T go to infinity, without any constraints on the relative speed of growth.

In contemporary economic applications, the PCA is often used to estimate factors F and loadings Λ in the factor model for the temporarily demeaned data[2]

$$X - \bar{X} = \Lambda F' + e. \tag{2}$$

The common factors are often interpreted as a few important latent variables affecting a vast number of economic indicators (rows of X). See Stock and Watson (2016) for a review of the related literature. Of course, in general, data generated from (1) do not have a factor structure. For example, if $\Psi(L)$ is diagonal, then the data are cross-sectionally independent and there are clearly no common factors.

Suppose that a researcher, nevertheless, models the data by (2). The PCA estimates of the first r factors are then defined as the r principal eigenvectors, $\hat{F}_1, ..., \hat{F}_r$, of

$$\hat{\Sigma} = \left(X - \bar{X}\right)' \left(X - \bar{X}\right) / N. \tag{3}$$

The corresponding principal eigenvalues $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_r$ estimate the explanatory power of the factors. Precisely, $\hat{\lambda}_j / \operatorname{tr} \hat{\Sigma}$ is interpreted as the fraction of the data variation explained by the j-th factor.

Below, we show that such a principal components analysis may be spurious in the sense that $\hat{\Sigma}$ has a few eigenvalues that dominate the rest, but the corresponding eigenvectors do not represent any latent economic factors driving the dynamics of the data. Instead, they capture deterministic trends that explain a large share of variation in any time series that are integrated of order one.

Denote the i-th component of the vector ε_t as ε_{it}, and let Ψ_k be the coefficients of the matrix lag polynomial $\Psi(L) = \sum_{k=0}^{\infty} \Psi_k L^k$. We make the following assumptions.

Assumption A1. *Random variables ε_{it} with $i \in \mathbb{N}$ and $t \in \mathbb{Z}$ are independent and such that $\mathbb{E}\varepsilon_{it} = 0$, $\mathbb{E}\varepsilon_{it}^2 = 1$, and $\varkappa_4 = \sup_{i \in \mathbb{N}, t \in \mathbb{Z}} \mathbb{E}\varepsilon_{it}^4 < \infty$.*

Note that ε_{it} may have different distributions, although they have to be independent. Further, the normalization $\mathbb{E}\varepsilon_{it}^2 = 1$ is not restrictive because it may be

[2] We consider the case of demeaned and standardized data in the next section.

accommodated by the lag polynomial $\Psi(L)$.

Assumption A2. *As* $N \to \infty$, $\sum_{k=0}^{\infty}(1+k)\|\Psi_k\| = O(N^{\alpha})$ *for some* $\alpha \geq 0$, *where* $\|\cdot\|$ *denotes the spectral norm.*

This assumption mildly restricts the form of temporal and cross-sectional dependence in the data. Although our setting *does not imply* the existence of common factors in the data, it does allow for them when $\alpha > 0$. For a simple example, consider a basic factor model

$$X_t = \Lambda F_t + e_t, \tag{4}$$

where the factors follow independent random walks, and the idiosyncratic component is white noise. As shown in the Appendix, such X_t can be represented in the form (1), where A2 holds with $\alpha = 1/2$.

Assumption A3. *As* $N \to \infty$, $\operatorname{tr}\Omega / \|\Omega\| \to \infty$, *where* $\Omega = \Psi(1)\Psi(1)'$ *is the long-run covariance matrix.*

Since $\operatorname{tr}\Omega / \|\Omega\| \leq \operatorname{rank}\Omega$, assumption A3 implies that the rank of $\Psi(1)$ diverges to infinity as $N \to \infty$. In other words, the number of stochastic trends in the data is increasing with the dimensionality.

The assumption does not allow a finite number of such trends to dominate the rest, so that $\|\Omega\|$ is not allowed to dominate $\operatorname{tr}\Omega$ asymptotically. In particular, A3 precludes the existence of a fixed number of strong nonstationary factors in the data. However, the existence of a growing number of such factors, a fixed number of weaker factors, or the total absence of any factors is allowed. We illustrate this with simple examples in the Appendix.

Theorem 1 *Let "*$\xrightarrow{\mathrm{P}}$*" denote convergence in probability. Suppose A1-A3 hold. If*

$$N^{2\alpha}(T + N_{\varepsilon}) / \left(T^2 \operatorname{tr}\Omega\right) \to 0 \tag{5}$$

as $N, T \to \infty$, *then for any fixed positive integer* k,
(i) $\left|\hat{F}_k' d_k\right| \xrightarrow{\mathrm{P}} 1$, *where* $d_k = (d_{k1}, ..., d_{kT})'$ *with* $d_{kt} = \sqrt{2/T}\cos(\pi k t/T)$.
(ii) $\hat{\lambda}_k / (\gamma_N T^2) \xrightarrow{\mathrm{P}} (k\pi)^{-2}$, *where* $\gamma_N = \operatorname{tr}\Omega / N$.

$$\textit{If } \min\{N, T\} N^{2\alpha}(T + N_{\varepsilon}) / \left(T^2 \operatorname{tr}\Omega\right) \to 0, \textit{ then} \tag{6}$$

(iii) $\hat{\lambda}_k / \operatorname{tr}\hat{\Sigma} \xrightarrow{\mathrm{P}} 6/(k\pi)^2$.

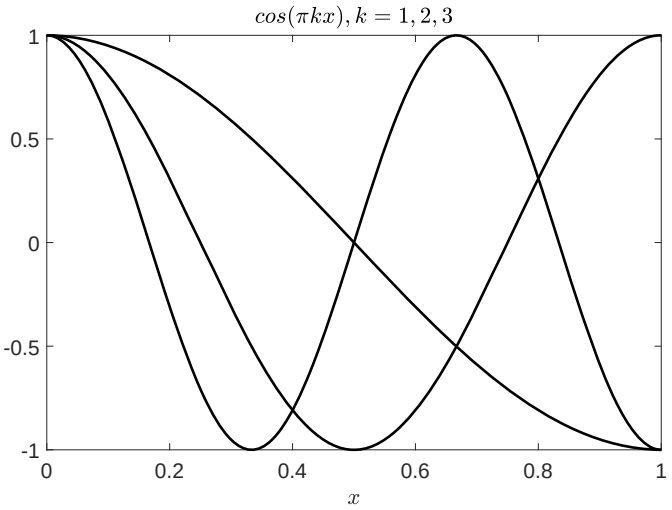

Figure 2: The probability "limits" of the first three spurious factor estimates.

Let us first interpret the theorem's results, and then discuss conditions (5) and (6) that link N, N_ε, T, and $\operatorname{tr}\Omega$.

Part (i) of the theorem reveals that the "factor" estimates converge in probability to deterministic cosine functions in the sense that the angle between the vector of estimates and the vector of uniform grid values of the corresponding cosine function converges in probability to zero. Figure 2 plots the cosine functions corresponding to the first three "factors". They may be interpreted as the trigonometric versions of the linear, quadratic, and cubic trends.

As mentioned in the Introduction, the functions can be linked to the Karhunen-Loève expansion of the demeaned Wiener process $\tilde{W}(x) = W(x) - \int_0^1 W(x)\,\mathrm{d}x$. Its covariance kernel has eigenfunctions $\sqrt{2}\cos(\pi kx)$, $k = 1, 2, ...$, corresponding to eigenvalues $(\pi k)^{-2}$ (e.g. Müller and Watson (2008, Thm. 1)). Therefore, the Karhunen-Loève expansion of $\tilde{W}(x)$ has the following form

$$\tilde{W}(x) = \sqrt{2}\sum_{k=1}^{\infty}(\pi k)^{-1}\cos(\pi kx)\, z_k, \tag{7}$$

where z_k are i.i.d. standard normal random variables.

For each of the data series X_{jt} that are difference-stationary, define $Y_{jT}(x) = (f_j(0)\,T)^{-1/2} X_{j[xT]}$, where $f_j(0)$ is the spectral density of $X_{jt} - X_{j,t-1}$ at frequency zero. As is well-known (e.g. Phillips, 1986), functions $Y_{jT}(x)$ weakly converge to

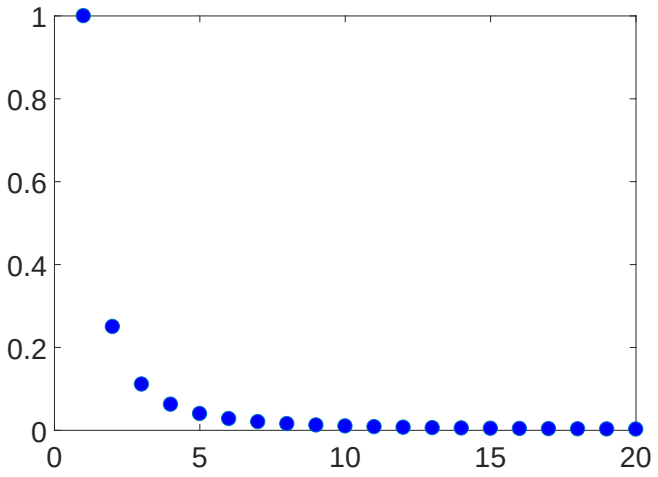

Figure 3: The asymptotic scree plot for possibly factorless persistent data (the first 20 normalized eigenvalues only). The horizontal axis shows the order k of the eigenvalue λ_k. The vertical axis shows the probability limit of λ_k/λ_1.

$W(x)$ and thus,

$$Y_{jT}(x) - \bar{Y}_{jT} = (f_j(0)\,T)^{-1/2}\left(X_{j[xT]} - \bar{X}_j\right)$$

weakly converge to $\tilde{W}(x)$. Therefore, the demeaned series X_{jt}, divided by $(f_j(0)\,T)^{1/2}$, can asymptotically be represented by the Karhunen-Loève expansion of $\tilde{W}(x)$. In particular, functions $\cos(\pi kt/T)$ with $k = 1, 2, ...$ capture much of the variation in each of $X_{jt} - \bar{X}_j$, which agrees with the theorem's first result intuitively.

The above arguments suggest that we should expect a flavour of spurious factor analysis to be present even in the PCA of the nonstationary data of fixed dimension N. The cosines would still be capturing much of the common variation in the data (regressing the data on them would produce high R^2), although the PCA estimators of the "factors" would no longer converge to these cosines.

Figure 3 illustrates statement (ii) of the theorem by showing the asymptotic scree plot for data satisfying the theorem's assumptions. The height of the plot is scaled so that the largest eigenvalue equals one. A typical interpretation of such a plot would be that the data "obviously" contain at least one strong factor, but perhaps two, or even three of them. Theorem 1 (ii) shows that such an interpretation may potentially be very misleading as the data may be totally factorless.

Part (iii) of the theorem describes the portion of data variation attributed to the

k-th principal component. A naive but standard interpretation of this result would be that the first k factors explain $\sum_{j=1}^{k} 6/(j\pi)^2 \times 100\%$ of the variation in the data. This "explanatory power" is very strong. The first three spurious factors absorb more than 80% of the data variation.

Let us now discuss conditions of the theorem that link N, N_ε, T, and $\operatorname{tr}\Omega$. For an extreme example where $\Psi(L) = I_N$, so the data consist of independent pure random walks, we have $\alpha = 0$, $N_\varepsilon = N$, and $\operatorname{tr}\Omega = N$. Hence, conditions (5) and (6) are trivially satisfied. It is easy to see that the conditions continue to hold for non-diagonal $\Psi(L)$ (so the data consist of cross-sectionally dependent I(1) processes) as long as $\Psi(L)$ satisfies A2 with $\alpha = 0$, $\|\Psi(1)\|$ and $\|\Psi(1)^{-1}\|$ remain bounded, and $N_\varepsilon = N$.

For example, let the first differenced data follow an autoregression $\Delta X_t = \rho \Delta X_{t-1} + e_t$ with $|\rho| < 1$, where $e_t = (e_{1t}, ..., e_{Nt})'$ are generated by "cross-sectional autoregressions" $e_{it} = \gamma e_{i-1,t} + \varepsilon_{it}$ with $|\gamma| < 1$ and $e_{0t} = 0$. Then $\Psi(L) = (1 - \rho L)^{-1} \Gamma_N$, where Γ_N is an N-dimensional lower triangular Toeplitz matrix with ones on the main diagonal and γ^j on the j-th sub-diagonal. As is well known (e.g. Böttcher and Silbermann (1999, Corollary 4.19)),

$$\lim_{N\to\infty} \|\Gamma_N\| = (1-\gamma)^{-1} \text{ and } \lim_{N\to\infty} \left\|\Gamma_N^{-1}\right\| = (1+\gamma)^{-1}.$$

Therefore, $\Psi(L)$ satisfies A2 with $\alpha = 0$, whereas $\|\Psi(1)\|$ and $\left\|\Psi(1)^{-1}\right\|$ converge to finite positive numbers.

Note that the conditions of the theorem do not require all data to be integrated. Suppose, for example, that $\Psi(L)$ is diagonal with first n diagonal elements equal one, and the rest equal $1 - L$. Then, the first n data series are random walks whereas the last $N - n$ series are white noise. Obviously, $N_\varepsilon = N$, A2 holds with $\alpha = 0$, and A3 holds when $\operatorname{tr}\Omega = n \to \infty$. Condition (5) becomes equivalent to $(T + N)/(T^2 n) \to 0$. Hence, *for any* $n \to \infty$, it holds with $N = O(T^2)$, whereas (6) holds with $N = O(T)$.

The fact that a relatively small number n of I(1) series so strongly influence the PCA results can be partially blamed on the different scale of I(1) and I(0) series. The effect of such a scale difference would be eliminated by standardizing the data. We study consequences of the standardization in the next section.

Here, we point out that the effect of the scale difference can also be eliminated by dividing I(1) series by $\sqrt{T}$. Such an adjustment transforms (5) to $(T + N)/(Tn) \to 0$.

For $N = O(T)$, this constraint is not binding under the maintained assumption that $n \to \infty$. In particular, if the data contain any increasing number of I(1) series, the PCA estimate of the first "factor" would converge to a deterministic cosine wave, *even after dividing* the I(1) series by $\sqrt{T}$.

Finally, consider the basic factor model example (4) with the number of factors $N_F \to \infty$. In that example, sufficient condition (5) for statements (i-ii) of the theorem to hold becomes $(T + N + N_F) / (T^2 N_F) \to 0$. In particular, if $N = O(T^2)$, the PCA estimates of a few of the strongest factors converge to deterministic cosine waves even though the data do contain an increasing number of genuine strong factors, which may be different from the cosine waves.

3 Extensions

3.1 Local level model

Suppose that data $Y_t, t = 1, ..., T$, are weighted sum of $I(1)$ and $I(0)$ components

$$Y_t = \omega_T X_t + Z_t, \tag{8}$$

where $\omega_T \neq 0$ is possibly decreasing with the sample size T, X_t is generated by the integrated system (1) as in the previous section, and Z_t is an N-dimensional linear stationary process. Specifically, $Z_t = \Pi(L)\eta_t$, where $\Pi(L) = \sum_{k=0}^{\infty} \Pi_k L^k$ and η_t is an N_η-dimensional random vector with components η_{it}. We make the following assumption.

Assumption A4. *Random variables η_{it} with $i \in \mathbb{N}$ and $t \in \mathbb{Z}$ are independent and such that $\mathbb{E}\eta_{it} = 0$, $\mathbb{E}\eta_{it}^2 = 1$, and $\tau_4 = \sup_{i\in\mathbb{N},t\in\mathbb{Z}} \mathbb{E}\eta_{it}^4 < \infty$. Further, $\sum_{k=0}^{\infty} (1 + k) \|\Pi_k\| = O\left(N^\beta\right)$ for some $\beta \geq 0$ as $N \to \infty$.*

The part of the assumption describing properties of η_{it} parallels assumption A1 for ε_{it}. We do not assume that η_{it} and ε_{it} are mutually independent so Z_t and X_t may depend on each other. The second part of A4 parallels A2. The constant β is introduced to allow component Z_t of the data to contain some genuine common factors.

Let $Y = [Y_1, ..., Y_T]$ be the $N \times T$ data matrix. Let $\check{\lambda}_1 \geq ... \geq \check{\lambda}_T$ and $\check{F}_1, ..., \check{F}_T$ be the eigenvalues and corresponding eigenvectors of $\check{\Sigma} = \left(Y - \bar{Y}\right)' \left(Y - \bar{Y}\right) / N$.

Theorem 2 *Under A1-A4, if (5) holds and*

$$N^{2\beta}\left(T+N_\eta\right)/\left(\omega_T^2T^2\operatorname{tr}\Omega\right)\to 0 \tag{9}$$

as $N,T\to\infty$, then for any fixed positive integer k,
(i) $\left|\check{F}_k'd_k\right|\xrightarrow{\mathrm{P}}1$, where $d_k=(d_{k1},...,d_{kT})'$ with $d_{kt}=\sqrt{2/T}\cos\left(\pi kt/T\right)$.
(ii) $\check{\lambda}_k/\left(\omega_T^2\gamma_NT^2\right)\xrightarrow{\mathrm{P}}(k\pi)^{-2}$, where $\gamma_N=\operatorname{tr}\Omega/N$.
(iii) If (6) holds and

$$\min\left\{N,T\right\}N^{2\beta}\left(T+N_\eta\right)/\left(\omega_T^2T^2\operatorname{tr}\Omega\right)\to 0, \tag{10}$$

then $\check{\lambda}_k/\operatorname{tr}\check{\Sigma}\xrightarrow{\mathrm{P}}6/\left(k\pi\right)^2$.

As an illustration, consider a simple situation where the components of X_t and Z_t are independent random walks and white noises, respectively. Then A2 holds with $\alpha=0$ and A4 holds with $\beta=0$. Furthermore, $\operatorname{tr}\Omega=N_\varepsilon=N_\eta=N$. Therefore, (5) trivially holds, whereas (9) holds if $\left(T+N\right)/\left(\omega_T^2T^2N\right)\to 0$. If T and N diverge to infinity proportionally, the latter convergence holds as long as ω_T goes to zero slower[3] than $1/T$.

For another example, suppose that the components of X_t are independent random walks, as in the previous example. However, Z_t now contains a strong stationary factor so that $Z_t=\Gamma f_t+e_t$, where $\Gamma'\Gamma=N$ while the factor f_t and the components of e_t are independent white noises. Then $\beta=1/2$ and $N_\eta=N+1$. Hence, for (9) to hold we need $N\left(T+N+1\right)/\left(\omega_T^2T^2N\right)\to 0$. If T and N are proportional, the latter convergence holds as long as ω_T goes to zero slower than $1/\sqrt{T}$.

Condition (10) is harder to satisfy than (9). In the first of the above examples, it requires that ω_T goes to zero slower than $1/\sqrt{T}$ (assuming that T and N are proportional). For the second example, it fails for ω_T that converges to zero at any rate when T and N are proportional, but holds for ω_T going to zero slower than $\sqrt{N/T}$ when T grows faster than N.

In a related paper, Onatski and Wang (2020), we consider data having local-to-unit roots and establish a theorem similar, in spirit, to Theorems 1 and 2.

[3]When $\omega_T=w/T$ with fixed $w>0$, the eigenvalues of the sample covariance matrix still decay very fast, although the probability limits described by Theorem 2 (ii) should be altered. Similarly, the eigenvectors become imperfectly collinear with the cosine waves described by (i). The interested reader can find a partial analysis of the situation $\omega_T=w/T$ with fixed $w>0$ in the SM's Section 3.1.2.

3.2 Demeaned and standardized data

In PCA applications, the data are often not only demeaned, but also standardized. As we show below, the spurious factor phenomenon is still present after the standardization.

We consider data generated by equation $X_t = X_{t-1} + \Psi(L)\varepsilon_t$, as in our basic setting. However, this time matrix $\hat{\Sigma}$ is defined as

$$\hat{\Sigma} = \left(X - \bar{X}\right)' D^{-1} \left(X - \bar{X}\right)/N,$$

where $D = \text{diag}\left\{\left(X - \bar{X}\right)\left(X - \bar{X}\right)'/T\right\}$. This change substantially complicates our technical analysis. It requires us working with high-dimensional matrices whose entries are ratios of quadratic forms instead of just quadratic forms. As a result, our proofs for the demeaned case do not go through.

To overcome the technical challenge we simplify our setting.

Assumption A2a. *Matrix lag polynomial $\Psi(L)$ is diagonal. There exist absolute constants $B > 0$ and $b > 0$ such that, for all N,*

$$\max_i \sum_{k=0}^{\infty} (1+k) \left|(\Psi_k)_{ii}\right| \leq B \text{ and } \min_i \left|\sum_{k=0}^{\infty} (\Psi_k)_{ii}\right| \geq b.$$

Most important, we now require $\Psi(L)$ be diagonal, so our data are cross-sectionally independent. Although cross-sectionally independent data are rare in PCA applications, they are clearly factorless. Our point is to show that the PCA of such factorless data yields spurious factors even after the data are standardized. We leave analysis of cross-sectionally dependent standardized data for future research.

The existence of B, described in A2a, would follow from the diagonality of $\Psi(L)$ and A2 with $\alpha = 0$. The existence of b, described in A2a, is assumed to further simplify our proofs. It implies that all data series are integrated, so that no series, when first differenced, have zero spectral density at zero frequency.

Theorem 3 *Suppose that assumptions A1, A2a, and A3 hold. In addition, suppose that ε_{jt} are identically distributed. Then, for any fixed positive integer k,*
(i) $\left|\hat{F}_k' d_k\right| \xrightarrow{\text{P}} 1$, *where* $d_k = (d_{k1}, ..., d_{kT})'$ *with* $d_{kt} = \sqrt{2/T}\cos(\pi kt/T)$.
(ii) $\hat{\lambda}_k/T \xrightarrow{\text{P}} \nu_k$, *where* $\nu_k = \mathbb{E}\left(z_k^2 / \sum_{j=1}^{\infty} (kz_j/j)^2\right)$ *with* z_j, $j = 1, 2, 3, ...$ *being i.i.d. standard normal random variables.*

(iii) $\hat{\lambda}_k / \operatorname{tr} \hat{\Sigma} \xrightarrow{\mathrm{P}} \nu_k$.

Part (i) of the theorem shows that the standardization does not affect the asymptotic behavior of the spurious factors. They still converge to the cosine waves. However, the standardization does affect the form of the normalization of $\hat{\lambda}_k$ in (ii), as well as the form of the limits in (ii) and (iii).

The standardization removes the need for normalizing $\hat{\lambda}_k$ by the average long run variance parameter $\gamma_N = \operatorname{tr} \Omega / N$, as in Theorem 1 (ii). Further, since the conditional variance of an integrated process is of order T, the standardization leads to the situation where $\hat{\lambda}_k$ in Theorem 3 (ii) is divided by T as opposed to T^2 in Theorem 1 (ii).

Note that the limit $6/(k\pi)^2$ in Theorem 1 (iii) can be written in the form $1/\sum_{j=1}^{\infty} (k/j)^2$. Therefore, this limit can be obtained from the limit ν_k in Theorem 3 (iii) by replacing the chi-square variables z_i^2 by their expectation (unity).

Values of ν_k for different k can be obtained numerically. Our calculations show that $\nu_1 \approx 0.44$, $\nu_2 \approx 0.18$, and $\nu_3 \approx 0.095$. Hence, the "explanatory power" of the first spurious factor in the standardized setting is substantially lower than that in the non-standardized one, $6/\pi^2 \approx 0.61$. However, the "explanatory power" of the second and third spurious factors somewhat increase relative to the non-standardized $6/(2\pi)^2 \approx 0.15$ and $6/(3\pi)^2 \approx 0.068$. Overall, the first three spurious factors still "explain" a very high 71.5% portion of variation in the factorless standardized data.

4 The "number of factors"

Now we return to the basic setup of Section 2 and ask the following question. What is the number of "factors" in factorless persistent data detected by information criteria? Bai (2004) proposes to estimate the number of factors in nonstationary panels by minimizing function

$$IPC(k) = V(k) + k\hat{\sigma}^2 p(N,T)$$

over $k = 0, 1, ..., k_{\max}$, where $V(k) = \operatorname{tr}\hat{\Sigma}/T - \sum_{j=1}^{k} \hat{\lambda}_j/T$, $\hat{\sigma}^2 = V(k_{\max})$, and $p(N,T)$ is one of the following three penalty functions

$$\begin{aligned} p_1(N,T) &= \alpha_T \frac{N+T}{NT} \log \frac{NT}{N+T}, \\ p_2(N,T) &= \alpha_T \frac{N+T}{NT} \log \delta_{NT}, \text{ or} \\ p_3(N,T) &= \alpha_T \frac{N+T-k}{NT} \log NT. \end{aligned}$$

Here $\alpha_T = T/(4 \log\log T)$, and $\delta_{NT} = \min\{N, T\}$.

Let us denote the value k that delivers the minimum of $IPC(k)$ based on penalty $p_j(N,T)$ as $\hat{k}_j$. Bai's (2004) Theorem 1 gives conditions under which $\hat{k}_j$ is consistent for the true number of factors. One of the theorem's assumptions is the weak temporary dependence of the idiosyncratic terms. Of course, it does not generally hold for data generated by integrated system (1). However, in actual empirical research, one would not know the validity of the assumptions. If the data are nonstationary, it would be natural to apply an IPC criterion.

As the following proposition shows, the asymptotic behavior of $\hat{k}_j$ is sensitive to the choice of $k_{\max}$. We consider the following two rules for choosing $k_{\max}$. One rule is fixing $k_{\max}$ independent of the data size. The other sets $k_{\max}$ at some small fraction of $\delta_{NT} = \min\{N, T\}$, say $k_{\max} = [\gamma \delta_{NT}]$.

Proposition 4 *Suppose A1-A3 and condition (6) of Theorem 1 hold. Further, let*

$$m_{NT} = \frac{1}{\delta_{NT}} + \frac{(T + N_\varepsilon) N^{2\alpha} \delta_{NT}}{T^2 \operatorname{tr}\Omega}.$$

(i) if $k_{\max}$ is fixed, then $\hat{k}_j \xrightarrow{\mathrm{P}} 0$ as $N, T \to \infty$, for $j = 1, 2, 3$;
(ii) if $k_{\max} = [\gamma \delta_{NT}]$ with fixed $\gamma > 0$ and $m_{NT} p_j(N,T) \to 0$, then $\hat{k}_j \xrightarrow{\mathrm{P}} \infty$ as $N, T \to \infty$, for $j = 1, 2, 3$.

For cases where N, N_ε, and T are of the same order of magnitude and $\alpha = 0$, the convergence $m_{NT} p_j(N,T) \to 0$ required by Proposition 4 (ii) is guaranteed if $p_j(N,T) / \operatorname{tr}\Omega \to 0$. The latter convergence holds whenever $\log N$ is asymptotically dominated by $\operatorname{tr}\Omega \log\log N$. This would happen, for example, for data that consist of i.i.d. random walks. Moreover, $\log N$ would be asymptotically dominated by

$\operatorname{tr}\Omega \log\log N$ even if the number of the random walks is $\log N$ while the rest $N - \log N$ series are white noises.

The strong sensitivity of IPC to the choice of $k_{\max}$ can be circumvented by the use of the logarithmic criteria of the form $\log V(k) + kg(N,T)$. In contrast to IPC, the logarithmic criteria do not have the scaling factor $\hat{\sigma}^2$ in the penalty, which therefore does not depend on $k_{\max}$. Bai (2004) shows the consistency of the corresponding $\hat{k}_{\log}$ under his assumptions (not holding in our setting) and when $g(N,T) \to \infty$ while $g(N,T)/\log T \to 0$. Unfortunately, since for any fixed k, $\log V(k) = O_{\mathrm{P}}\left(\log \frac{T \operatorname{tr}\Omega}{N}\right)$, we immediately see that penalties satisfying the latter requirement yield $\hat{k}_{\log} \xrightarrow{\mathrm{P}} \infty$ as long as $\operatorname{tr}\Omega/N$ remains bounded away from zero.

In the Appendix we perform a Monte Carlo analysis of the finite sample behavior of $\hat{k}_j$ when data do not have any factors in them. We find that for empirically relevant data sizes and standard choices of $k_{\max}$, the estimated number of "factors" often equals two or three.

5 Problem detection

As we have seen above, factor analysis applied directly to large nonstationary panels may be spurious. This raises a question: how to detect spurious results? A simple, although inexact, check is to compare the time series plots of the estimated factors to the cosine functions. A similarity should raise the alarm.

As an example, consider Bai's (2004) analysis of sectoral employment in the US. Figure 4 replicates Figure 3 in Bai (2004). It shows the Bureau of Economic Analysis data (NIPA, Tables 6.5b and 6.5c) for the logarithm of employment across 58 sectors[4] in the US for the period from 1948 to 2000. The series are very persistent, and Bai (2004) identifies two nonstationary and one stationary factors in the data.

Figure 5 shows the time series plots of the PCA estimates of the three factors. Their resemblance to cosine functions is striking. It suggests that an extra caution should be exercised before structural interpretation of these factor estimates is attempted.

A more formal problem detection strategy consists of comparing factor estimates from the data in levels to those from the differenced data. If all the nonstationarity

[4]Bai (2004) has data on 60 sectors. However, the data on two out of 60 sectors in the current versions of NIPA tables is incomplete. Therefore, we use 58 sectors.

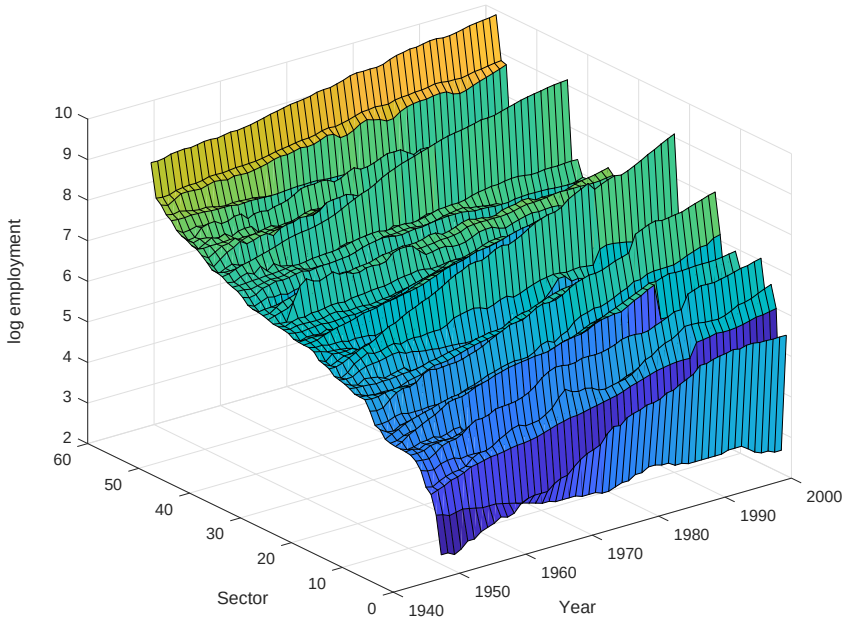

Figure 4: The number of full-time equivalent employees across 58 sectors. The sectors are arranged in ascending order according to their 1948 values.

in the data comes from factors, then under assumptions of Bai (2004) the PCA estimates $\hat{F}$ are consistent (up to a non-degenerate linear transformation) for the true factors F. Similarly, under assumptions of Bai and Ng (2004), the estimates $\hat{f}$ of the factors in the differenced data are consistent for ΔF. In such a case, $\Delta\hat{F}$ should be well aligned with $\hat{f}$. In contrast, a poor alignment would signal spurious results. A rigorous implementation of this strategy requires a separate research effort which we currently undertake.

6 Conclusion

This paper warns empirical researchers that a very high explanatory power of a few principal components of nonstationary data does not necessarily indicate the presence of factors. Even if such data are cross-sectionally independent, the first k principal components must explain $\sum_{j=1}^{k} 6/\left(j\pi\right)^2 \times 100\%$ of the variation, asymptotically. The extracted spurious factors correspond to the eigenfunctions of the auto-covariance

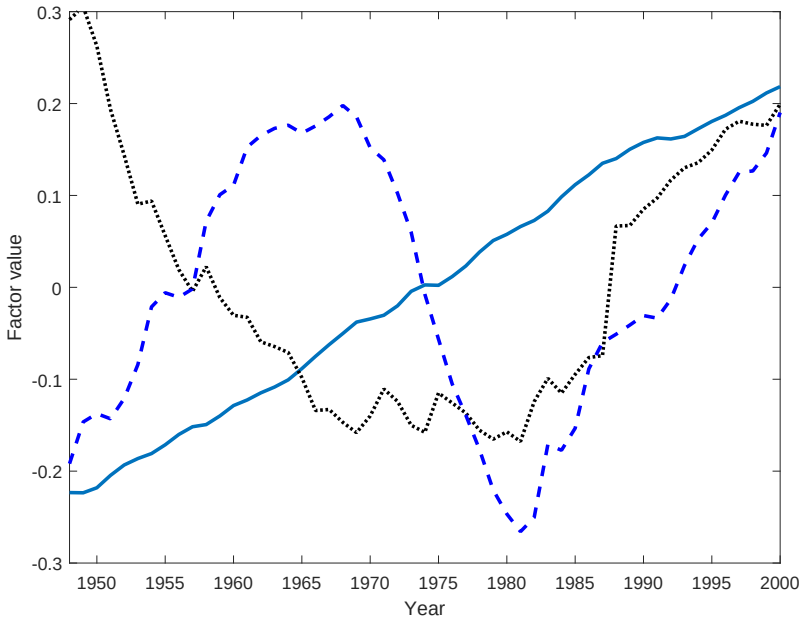

Figure 5: The principal components estimates of three factors in the employment data. The estimates are normalized to have unit Euclidean norms.

kernel of the Wiener process and do not represent any cross-sectional common shocks driving the data's dynamics.

Unfortunately, the standard criteria for the determination of the number of factors are sensitive to the choice of the maximum number of factors $k_{\max}$. For empirically relevant data sizes and standard choices of $k_{\max}$, such criteria would often suggest two or three factors, when in fact, none are present. Moreover, checking the stationarity of the PCA residuals using the Dickey-Fuller tests may spuriously favour the stationarity hypothesis. This may mislead a researcher to conclude that all the non-stationarity in the data is captured by a few common factors, which are consistently estimated by the PCA.

To detect these potential problems, we propose to always look at the time series plots of the extracted factors. Their resemblance to cosine waves should raise the alarm. A more formal detection strategy would compare the factor estimates obtained from the data in levels and in first differences.

Mis-interpreting spurious factors as common shocks driving economic data may mislead structural economic analysis. Less obvious, using such factors in forecasting

exercises may lead to forecast sub-optimality.[5] This can be clearly seen in the extreme situation where all data are independent random walks. For such data, optimal forecasts equal the most recent observations. They would be different from the forecasts based on the cosine waves that represent the spurious factors asymptotically.

In conclusion, we would like to stress that our critique does not apply to *all* PC analysis in economics. Most of this analysis is careful with respect to the assumptions made and is, therefore, immune to our critique. Furthermore, we would be very disappointed if some readers conclude from our analysis that there are no common economic forces affecting various economic data series. In the literature, there is ample evidence that such common forces are often present, which gives an indisputable value to careful economic research based on high-dimensional factor analysis.

7 Appendix

7.1 Discussion of A2 and A3

Let us show that A2 allows for factors in the data. Consider N-dimensional data X_t that satisfy factor model (4) with N_F factors that follow independent random walks. Assume that the loadings are normalized so that $\Lambda'\Lambda/N = I_{N_F}$, and the idiosyncratic component is white noise.

Such X_t satisfies (1) with ε_t that stacks vectors $F_t - F_{t-1}$ and e_t, and a linear lag polynomial $\Psi(L)$ with matrix coefficients $\Psi_0 = [\Lambda, I_N]$ and $\Psi_1 = -[0, I_N]$. We have

$$\|\Psi_0\| = \|\Psi_0\Psi_0'\|^{1/2} = (\|\Lambda\Lambda'\| + 1)^{1/2} = \sqrt{N+1},$$

and $\|\Psi_1\| = 1$. Therefore, $\sum_{k=0}^{\infty}(1+k)\|\Psi_k\| = \sqrt{N+1}+2$, and hence A2 is satisfied with $\alpha = 1/2$.

This simple example also illustrates the fact that the existence of a growing number of strong non-stationary factors in the data is allowed by A3. Indeed, the N_F factors in X_t are non-stationary and strong in the sense that sums of their squared loadings are proportional (equal in this example) to the sample size. We have $\Psi(1) = [\Lambda, 0]$ and $\Lambda'\Lambda/N = I_{N_F}$. Therefore, $\operatorname{tr}\Omega = \operatorname{tr}(\Psi(1)\Psi(1)') = NN_F$ and $\|\Omega\| = N$. We see that $\operatorname{tr}\Omega/\|\Omega\| = N_F$ and assumption A3 is indeed satisfied as long N_F is growing.

[5] We are grateful to James Stock for pointing out this fact to us.

(N,T)	Content	Source
$(60,52)$	US annual industry-level employment.	Bai (2004)
$(243,83)$	European quarterly macroeconomic data	Boivin et al. (2009)
$(128,710)$	Current version of FRED-MD monthly macroeconomic dataset	McCracken and Ng (2015)
$(58,220)$	US quarterly "real activity dataset"	Stock and Watson (2016)

Table 1: The dimensionalities of datasets used in the analysis below.

The rate of this growth may be arbitrarily slow.

For another example, let the data consist of N independent pure random walks, so there are no common factors whatsoever. Then $\Psi(1) = I_N$ and $\operatorname{tr}\Omega / \|\Omega\| = N$. Hence A3 is again satisfied.

In recent studies of high-dimensional problems (e.g. Vershynin (2012), Koltchinskii and Lounici (2016)), the ratio of the trace of a matrix to its norm is called effective rank, or effective dimension. Large effective rank indicates that the matrix cannot be well approximated by a matrix of low rank. In context of A3, this means that the long-run covariance matrix does not have a low rank component that asymptotically dominates the remainder. This can be interpreted as the absence of a small number of strong non-stationary factors in the data.

We would like to stress that the failure of A3 in situations where data contain a fixed number of such strong factors does not depend on whether the idiosyncratic terms are stationary or not. For example, if e_t in (4) consists of independent random walks instead of white noises, X_t satisfies (1) with $\varepsilon_t' = \left(F_t' - F_{t-1}', e_t' - e_{t-1}'\right)$ and $\Psi(L) = [\Lambda, I_N]$. Hence, $\Omega = \Lambda\Lambda' + I_N$ and its effective rank equals $N(N_F+1)/(N+1)$, which remains bounded with fixed N_F, so that A3 is violated.

On the other hand, A3 still holds when the data contain a fixed number of weaker factors (such that $\|\Lambda\Lambda'\| = o(N)$). Then, the effective rank of $\Omega = \Lambda\Lambda' + I_N$ is no smaller than $N/(1+o(N))$, which obviously diverges as required by A3.

7.2 Monte Carlo analysis of the number of spurious factors

We simulate data on N i.i.d. Gaussian random walks of length T, where the (N,T)-pairs correspond to the dimensions of four actual datasets described in Table 1. The number of Monte Carlo (MC) replications is set to 10,000.

Table 2 reports the obtained MC distributions of $\hat{k} = \hat{k}_1$, the estimate of the number of factors produced by IPC with penalty $p_1(N,T)$. Results for $\hat{k}_2$ and $\hat{k}_3$ are similar and not reported. The columns of the table correspond to different choices of $k_{\max} = 6, ..., 15$. The entries of the table are the empirical probabilities (in percent rounded to the nearest integer) of observing a particular value of $\hat{k}$, which is given in the first column.

We see that the MC distributions of $\hat{k}$ concentrate at $\hat{k} = 2$ or $\hat{k} = 3$ for most of the settings. For example, when $k_{\max} = 10$ and $(N,T) = (60,52)$, the MC probability of observing $\hat{k} = 3$ equals 91%. For the same $k_{\max}$ and $(N,T) = (243,83)$, this probability becomes 100%. For $(N,T) = (128,710)$ and $(N,T) = (58,220)$, the mode of the MC distributions of $\hat{k}$ shifts to $\hat{k} = 1$ (probability 100%) and $\hat{k} = 2$ (probability 86%), respectively. Overall we see that, for empirically relevant data sizes, IPC criteria would typically estimate a small non-zero number of factors in the factorless persistent data.

7.3 Proof of Theorem 1

Consider the multivariate Beveridge-Nelson decomposition of the demeaned X_t

$$X_t - \bar{X} = \Psi(1)\left(\xi_t - \bar{\xi}\right) + \Psi^*(L)\left(\varepsilon_t - \bar{\varepsilon}\right),$$

where $\Psi^*(L) = \sum_{k=0}^{\infty} \Psi_k^* L^k$ with $\Psi_k^* = -\sum_{i=k+1}^{\infty} \Psi_i$, and $\xi_t = \sum_{j=1}^{t} \varepsilon_j$. In matrix notations,

$$XM = \Psi(1)\,\varepsilon UM + \Psi^*(L)\,\varepsilon M, \tag{11}$$

where M is the projection matrix on the space orthogonal to the T-dimensional vector of ones, ε is the $N_\varepsilon \times T$ matrix with columns ε_t, and U is the upper triangular matrix with ones above and on the main diagonal.

Recall that $\hat{\Sigma}$ is the sample covariance matrix of the demeaned data XM. Let $\tilde{\Sigma}$ be the sample covariance of the I(1) term in the Beveridge-Nelson decomposition (11), that is

$$\tilde{\Sigma} = MU'\varepsilon' W \varepsilon UM/N, \tag{12}$$

where $W = \Psi(1)'\Psi(1)$. Denote the eigenvalues of $\tilde{\Sigma}$ as $\tilde{\lambda}_1 \geq ... \geq \tilde{\lambda}_T$ and corresponding eigenvectors as $\tilde{F}_1, ..., \tilde{F}_T$. Since variation of I(1) series dominates that of I(0) series, it is reasonable to expect that $\hat{\Sigma}$ and $\tilde{\Sigma}$ are close in some sense. Therefore

k_{max}	6	7	8	9	10	11	12	13	14	15
					$(N,T) = (60,52)$ as in Bai (2004)					
$\hat{k} = 0$	0	0	0	0	0	0	0	0	0	0
$\hat{k} = 1$	0	0	0	0	0	0	0	0	0	0
$\hat{k} = 2$	96	75	43	15	4	1	0	0	0	0
$\hat{k} = 3$	4	25	57	84	91	79	54	27	9	2
$\hat{k} = 4$	0	0	0	1	5	20	46	72	85	77
$\hat{k} = 5$	0	0	0	0	0	0	0	1	6	21
					$(N,T) = (243,83)$ as in Boivin et al. (2009)					
$\hat{k} = 0$	0	0	0	0	0	0	0	0	0	0
$\hat{k} = 1$	0	0	0	0	0	0	0	0	0	0
$\hat{k} = 2$	99	76	23	2	0	0	0	0	0	0
$\hat{k} = 3$	1	24	77	98	100	98	84	50	19	4
$\hat{k} = 4$	0	0	0	0	0	2	16	50	81	96
					$(N,T) = (128,710)$ as in FRED-MD dataset, McCracken and Ng (2015)					
$\hat{k} = 0$	0	0	0	0	0	0	0	0	0	0
$\hat{k} = 1$	100	100	100	100	100	97	88	67	41	20
$\hat{k} = 2$	0	0	0	0	0	3	12	33	59	80
					$(N,T) = (58,220)$ as in "real activity dataset", Stock and Watson (2016)					
$\hat{k} = 0$	0	0	0	0	0	0	0	0	0	0
$\hat{k} = 1$	98	87	62	33	14	4	1	0	0	0
$\hat{k} = 2$	2	13	38	67	86	96	98	95	84	67
$\hat{k} = 3$	0	0	0	0	0	0	1	5	16	33

Table 2: The Monte Carlo distribution of the number of factors estimated using IPC_1 criterion. The probabilities in columns are measured in percent rounded to the nearest integer. The data are N independent random walks of length T each. The number of MC replications is 10,000.

our proof strategy is, first, show that Theorem 1 holds when $\hat{\lambda}_k, \hat{F}_k, \hat{\Sigma}$ are replaced by $\tilde{\lambda}_k, \tilde{F}_k, \tilde{\Sigma}$ and then, prove that replacing back "tildes" by "hats" does not affect the theorem's validity.

7.3.1 Proof of Theorem 1 for $\tilde{\lambda}_k, \tilde{F}_k, \tilde{\Sigma}$

First, we will prove the theorem for $k = 1$. Then, we handle general k by mathematical induction. The following technical lemma is established in SM.

Lemma 5 *Matrix MU' admits the singular value decomposition $MU' = \sum_{q=1}^{T} \sigma_q w_q v_q'$, where for $q < T$, $\sigma_q = (2\sin(\pi q/(2T)))^{-1}$ and the s-th coordinates of vectors w_q and v_q equal, respectively*

$$w_{qs} = -\sqrt{2/T}\cos((s-1/2)\pi q/T) \text{ and } v_{qs} = \sqrt{2/T}\sin((s-1)\pi q/T).$$

For $q = T$ we have $\sigma_T = 0$, $w_T = l_T/\sqrt{T}$, and $v_T = e_1$, where l_T is the T-dimensional vector of ones and e_1 is the first coordinate vector of $\mathbb{R}^T$.

Since w_q, $q = 1, ..., T-1$, form an orthonormal basis in the space orthogonal to l_T and $\tilde{F}_1$ belongs to this space, we have a representation

$$\tilde{F}_1 = \sum_{q=1}^{T-1} \alpha_q w_q. \tag{13}$$

Let us show that $\alpha_1^2 \xrightarrow{\mathrm{P}} 1$. This would establish part (i) of the theorem because $(w_1' d_1)^2 \to 1$.

Representation (13) and Lemma 5 yield

$$N\tilde{\lambda}_1 = N\sum_{r,q=1}^{T-1} \alpha_r \alpha_q w_r' \tilde{\Sigma} w_q = \gamma' W \gamma,$$

where $\gamma = \sum_{r=1}^{T-1} \alpha_r \sigma_r \varepsilon v_r$. The idea of the proof consists of, first, showing that the sum in the latter display is dominated by the terms $\alpha_r^2 w_r' \tilde{\Sigma} w_r$, and, then, demonstrating that $w_r' \tilde{\Sigma} w_r$ is quickly decreasing in r so that the maximum of the sum with respect to α's is achieved when α_1^2 is close to unity whereas α_r^2 with $r > 1$ are close to zero. We will need the following lemma. Its proof can be found in SM.

Lemma 6 *Suppose assumptions A1 and A3 hold. Then, for any positive integers i, j such that $i \le j \le T$,*

$$v_i' \varepsilon' W \varepsilon v_j = \operatorname{tr} W (\delta_{ij} + o_{\mathrm{P}}(1)), \tag{14}$$

where δ_{ij} is the Kronecker delta, and

$$\begin{aligned}\sum_{r=i}^{j}\sigma_r^2 v_r'\varepsilon' W\varepsilon v_r &= \operatorname{tr} W\sum_{r=i}^{j}\sigma_r^2\left(1+o_{\mathrm{P}}(1)\right) \\ &= \operatorname{tr} W\sum_{r=i}^{j}\sigma_r^2+\operatorname{tr} W o_{\mathrm{P}}(T^2).\end{aligned} \tag{15}$$

Let K be a fixed non-negative integer. Consider a decomposition $\gamma=\gamma_1+\gamma_2$, where

$$\gamma_1=\sum_{r=1}^{K}\alpha_r\sigma_r\varepsilon v_r \text{ and } \gamma_2=\sum_{r=K+1}^{T-1}\alpha_r\sigma_r\varepsilon v_r.$$

We have, by the triangle inequality,

$$\gamma' W\gamma\leq\left(\left(\gamma_1' W\gamma_1\right)^{1/2}+\left(\gamma_2' W\gamma_2\right)^{1/2}\right)^2. \tag{16}$$

Since K is fixed and $\sigma_r^2=O\left(T^2\right)$, equation (14) yields

$$\begin{aligned}\gamma_1' W\gamma_1 &= \operatorname{tr} W\sum_{r=1}^{K}\alpha_r^2\sigma_r^2+\operatorname{tr} W\, o_{\mathrm{P}}\left(T^2\right) \\ &\leq \operatorname{tr} W\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2+\operatorname{tr} W\, o_{\mathrm{P}}\left(T^2\right).\end{aligned} \tag{17}$$

Note that $\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2\leq\sigma_1^2=\sin^{-2}\left(\pi/\left(2T\right)\right)/4$. Since $\sin x\geq 2x/\pi$ for $x\in[0,\pi/2]$, we have $\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2\leq T^2/4$.

Further, we have

$$\gamma_2' W\gamma_2=\sum_{j=1}^{N_\varepsilon}\left(\sum_{r=K+1}^{T-1}\alpha_r\sigma_r\left[W^{1/2}\varepsilon v_r\right]_j\right)^2,$$

where $\left[W^{1/2}\varepsilon v_r\right]_j$ is the j-th component of vector $W^{1/2}\varepsilon v_r$. Recall that $\sum_{r=1}^{T-1}\alpha_r^2=1$. Therefore, by the Cauchy-Schwarz inequality,

$$\gamma_2' W\gamma_2\leq\sum_{j=1}^{N_\varepsilon}\sum_{r=K+1}^{T-1}\left(\sigma_r\left[W^{1/2}\varepsilon v_r\right]_j\right)^2=\sum_{r=K+1}^{T-1}\sigma_r^2 v_r'\varepsilon' W\varepsilon v_r.$$

This inequality and (15) yield

$$\gamma_2' W\gamma_2\leq\operatorname{tr} W\sum_{r=K+1}^{T-1}\sigma_r^2+\operatorname{tr} W o_{\mathrm{P}}(T^2). \tag{18}$$

Note that

$$\sum_{r=1}^{T-1}\sigma_r^2=\sum_{r=1}^{T-1}\left(4\sin^2\left(\pi r/\left(2T\right)\right)\right)^{-1}\leq\left(T^2/4\right)\sum_{r=1}^{T-1}r^{-2}. \tag{19}$$

Using (17) and (18) in (16), we obtain

$$\begin{aligned}\gamma'W\gamma &\leq \operatorname{tr}W\left(\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2+\sum_{r=K+1}^{T-1}\sigma_r^2\right. \\ &\quad\left.+2\left(\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2\sum_{r=K+1}^{T-1}\sigma_r^2\right)^{1/2}+o_{\mathrm{P}}\left(T^2\right)\right).\end{aligned} \tag{20}$$

Let us choose K so that $\sum_{r=K+1}^{T-1}\sigma_r^2\leq\delta^2T^2/4$, where $\delta<1$ is an arbitrarily small positive number. Such a choice should be possible in view of (19). Then, from (20),

$$\begin{aligned}\gamma'W\gamma &\leq \operatorname{tr}W\left(\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2+\left(\frac{1}{4}\delta^2+\frac{1}{2}\delta\right)T^2+o_{\mathrm{P}}\left(T^2\right)\right) \\ &\leq \operatorname{tr}W\left(\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2+\delta T^2+o_{\mathrm{P}}\left(T^2\right)\right).\end{aligned}$$

Since δ can be made arbitrarily small,

$$\gamma'W\gamma\leq\operatorname{tr}W\left(\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2+o_{\mathrm{P}}\left(T^2\right)\right).$$

Now recall that $\gamma'W\gamma=N\tilde{\lambda}_1$. Since

$$\sum_{r=1}^{T-1}\alpha_r^2\sigma_r^2\operatorname{tr}W\leq\alpha_1^2\sigma_1^2\operatorname{tr}W+(1-\alpha_1^2)\sigma_2^2\operatorname{tr}W,$$

we have

$$N\tilde{\lambda}_1\leq\alpha_1^2\sigma_1^2\operatorname{tr}W+(1-\alpha_1^2)\sigma_2^2\operatorname{tr}W+o_{\mathrm{P}}\left(T^2\right)\operatorname{tr}W. \tag{21}$$

On the other hand, $N\tilde{\lambda}_1$ must be no smaller than $Nw_1'\tilde{\Sigma}w_1=\sigma_1^2v_1'\varepsilon'W\varepsilon v_1$. By Lemma 6,

$$\sigma_1^2v_1'\varepsilon'W\varepsilon v_1=\sigma_1^2\operatorname{tr}W+o_{\mathrm{P}}\left(T^2\right)\operatorname{tr}W. \tag{22}$$

Therefore,

$$N\tilde{\lambda}_1\geq\sigma_1^2\operatorname{tr}W+o_{\mathrm{P}}\left(T^2\right)\operatorname{tr}W. \tag{23}$$

Combining this with (21), we obtain

$$\sigma_1^2 \operatorname{tr} W + o_{\mathrm{P}}(1) T^2 \operatorname{tr} W \leq \alpha_1^2 \sigma_1^2 \operatorname{tr} W + (1 - \alpha_1^2) \sigma_2^2 \operatorname{tr} W + o_{\mathrm{P}}(1) T^2 \operatorname{tr} W,$$

which yields

$$1 - \alpha_1^2 \leq o_{\mathrm{P}}(1) T^2 / \left(\sigma_1^2 - \sigma_2^2\right) = o_{\mathrm{P}}(1). \tag{24}$$

Hence,

$$\alpha_1^2 = \left(\tilde{F}_1' w_1\right)^2 \xrightarrow{\mathrm{P}} 1, \tag{25}$$

which completes our proof of statement (i) for $k = 1$.

To establish (ii), note that inequalities (21) and (23) yield

$$\left| N \tilde{\lambda}_1 - \sigma_1^2 \operatorname{tr} W \right| \leq \left| 1 - \alpha_1^2 \right| \left(\sigma_1^2 + \sigma_2^2\right) \operatorname{tr} W + o_{\mathrm{P}}\left(T^2\right) \operatorname{tr} W.$$

Combining this with the facts that $\alpha_1^2 = 1 + o_{\mathrm{P}}(1)$ and $\sigma_1^2 = T^2/\pi^2 + o(T^2)$, we obtain

$$\tilde{\lambda}_1 = \frac{T^2 \operatorname{tr} W}{\pi^2 N} (1 + o_{\mathrm{P}}(1)) = \frac{T^2 \operatorname{tr} \Omega}{\pi^2 N} (1 + o_{\mathrm{P}}(1)), \tag{26}$$

as claimed by statement (ii).

Further, by Lemma 5,

$$N \operatorname{tr} \tilde{\Sigma} = \operatorname{tr}\left(\sum\nolimits_{r=1}^{T} \sigma_r w_r v_r' \varepsilon' W \varepsilon \sum\nolimits_{q=1}^{T} \sigma_q v_q w_q'\right) = \sum\nolimits_{r=1}^{T} \sigma_r^2 v_r' \varepsilon' W \varepsilon v_r,$$

where the last equality follows from the orthonormality of the basis $\{w_r, r = 1, ..., T\}$. Hence, by Lemma 6,

$$N \operatorname{tr} \tilde{\Sigma} = \operatorname{tr} W \sum_{r=1}^{T} \sigma_r^2 (1 + o_{\mathrm{P}}(1)). \tag{27}$$

On the other hand, for any fixed K,

$$\sum_{r=1}^{K} \sigma_r^2 / T^2 = \sum_{r=1}^{K} 1 / \left(4T^2 \sin^2(\pi r / (2T))\right) \to \sum_{r=1}^{K} 1 / (\pi r)^2$$

as $T \to \infty$. Furthermore, $\sum_{r=K+1}^{T} \sigma_r^2 / T^2$ can be made arbitrarily small by choosing sufficiently large K. Hence, the Euler formula $\sum_{r=1}^{\infty} r^{-2} = \pi^2/6$ yields $\sum_{r=1}^{T} \sigma_r^2 / T^2 \to 1/6$.

The latter convergence and (27) give us

$$\operatorname{tr}\tilde{\Sigma} = \frac{T^2}{6N}\operatorname{tr} W\left(1 + o_{\mathrm{P}}\left(1\right)\right) = \frac{T^2}{6N}\operatorname{tr}\Omega\left(1 + o_{\mathrm{P}}\left(1\right)\right).$$

Combining this with (26), we obtain

$$\tilde{\lambda}_1 / \operatorname{tr}\tilde{\Sigma} = \left(6/\pi^2\right)\left(1 + o_{\mathrm{P}}(1)\right), \tag{28}$$

which concludes the proof of the theorem for $k = 1$.

For $k = m > 1$, the theorem follows by mathematical induction. Indeed, suppose it holds for $k < m$. Consider a representation $\tilde{F}_m = \sum_{q=1}^{T-1} \alpha_q w_q$. Since $\tilde{F}_m'\tilde{F}_j = 0$ for all $j < m$, and since $\left|\tilde{F}_j' w_j\right| = 1 + o_{\mathrm{P}}\left(1\right)$ by the induction hypothesis, we must have $\alpha_j = o_{\mathrm{P}}\left(1\right)$ for all $j < m$. In particular,

$$N\tilde{F}_m'\tilde{\Sigma}\tilde{F}_m = \sum\nolimits_{q,r=m}^{T-1} \alpha_q\alpha_r\sigma_q\sigma_r v_q'\varepsilon' W\varepsilon v_r + o_{\mathrm{P}}\left(T^2\right)\operatorname{tr} W. \tag{29}$$

To see that (29) holds, it is sufficient to establish equalities $N\alpha_j w_j'\tilde{\Sigma}\sum_{r=m}^{T-1}\alpha_r w_r = o_{\mathrm{P}}\left(T^2\right)\operatorname{tr} W$ for any $j < m$, and equalities $N\alpha_j\alpha_r w_j'\tilde{\Sigma}w_r = o_{\mathrm{P}}\left(T^2\right)\operatorname{tr} W$ for any $j, r < m$. Such equalities easily follow from the facts that $\alpha_j = o_{\mathrm{P}}\left(1\right)$ for all $j < m$ and $N\left\|\tilde{\Sigma}\right\| = N\tilde{\lambda}_1 = \left(T^2/\pi^2\right)\operatorname{tr} W\left(1 + o_{\mathrm{P}}(1)\right)$.

In addition to (29), we must have

$$N\sum_{i=1}^{m-1}\tilde{\lambda}_i + N\tilde{F}_m'\tilde{\Sigma}\tilde{F}_m \geq \sum_{i=1}^{m} w_i' MU'\varepsilon' W\varepsilon UMw_i = \left(\sum_{i=1}^{m}\sigma_i^2 + o_{\mathrm{P}}\left(T^2\right)\right)\operatorname{tr} W,$$

where the latter equality is obtained similarly to (22). Combining the above two displays, and using the induction hypothesis, this time regarding the validity of the identities $N\tilde{\lambda}_i = \left(\sigma_i^2 + o_{\mathrm{P}}\left(T^2\right)\right)\operatorname{tr} W$ for all $i < m$, we obtain

$$\sum_{q,r=m}^{T-1}\alpha_q\alpha_r\sigma_q\sigma_r v_q'\varepsilon' W\varepsilon v_r \geq \sigma_m^2\operatorname{tr} W + o_{\mathrm{P}}\left(T^2\right)\operatorname{tr} W. \tag{30}$$

Statements (i), (ii), and (iii) for $k = m$ now follow by arguments that are very similar to those used above for the case $k = 1$.

That is, we represent the sum on the left hand side of (30) in the form $\gamma' W \gamma$, where $\gamma = \sum_{r=m}^{T-1} \alpha_r \sigma_r \varepsilon v_r$. Then proceed along the lines of the above proof to obtain an upper bound on $\gamma' W \gamma$, similar to the right hand side of (21). Then, combining this upper bound with the lower bound (30), we prove the convergence $\alpha_m^2 \xrightarrow{\mathrm{P}} 1$. Finally, we proceed to establishing parts (ii) and (iii) using part (i). We omit details to save space.

7.3.2 Proof of Theorem 1 for $\hat{\lambda}_k, \hat{F}_k, \hat{\Sigma}$

We need to show that the theorem's validity for $\tilde{F}_k$, $\tilde{\lambda}_k$ and $\tilde{\Sigma}$ implies its validity for $\hat{F}_k$, $\hat{\lambda}_k$ and $\hat{\Sigma}$. By standard perturbation theory (e.g. Kato (1980), ch.2), such an implication for statements (i) and (ii) would follow if we are able to show that $\left\| \hat{\Sigma} - \tilde{\Sigma} \right\| = \frac{T^2}{N} \operatorname{tr} W \, o_{\mathrm{P}}(1)$. That is, the norm of $\hat{\Sigma} - \tilde{\Sigma}$ is asymptotically dominated by the sizes of the gaps between adjacent eigenvalues, $\tilde{\lambda}_k - \tilde{\lambda}_{k+1}$ and $\tilde{\lambda}_{k-1} - \tilde{\lambda}_k$. The Beveridge-Nelson decomposition (11) implies that it is sufficient to show that $\|\Psi^*(L)\, \varepsilon M\|^2 = T^2 \operatorname{tr} W \, o_{\mathrm{P}}(1)$. To establish this equality, we need the following lemma.

Lemma 7 *Suppose that assumption A1 holds. Let* $Z_t = \Pi(L)\varepsilon_t$ *and* $Z = [Z_1, ..., Z_T]$, *where* $\Pi(L) = \sum_{k=0}^{\infty} \Pi_k L^k$ *is an* $N \times N_\varepsilon$ *matrix lag polynomial that may depend on* N, N_ε, *and* T. *If* $\sum_{k=0}^{T} \|\Pi_k\| = O(N^\alpha)$ *and* $T \sum_{k=T+1}^{\infty} \|\Pi_k\|_F^2 = O(N_\varepsilon N^{2\alpha})$ *for an* $\alpha \geq 0$, *where* $\|\cdot\|_F$ *denotes the Frobenius norm, then*

$$\|Z\| = O_{\mathrm{P}}\left(T^{1/2} N^\alpha + N_\varepsilon^{1/2} N^\alpha\right). \tag{31}$$

We feel that the lemma might have an independent interest, especially for those researchers who assume that idiosyncratic terms of a factor model follow a linear process. Therefore, we provide a proof of the lemma below.

Proof: The lemma is a modification of Proposition 1 from Onatski (2015), where a proportional asymptotic regime with N/T converging to a nonzero constant is considered. The triangle inequality yields

$$\|Z\| \leq \sum_{k=0}^{T} \|\Pi_k\| \, \|\varepsilon_{-k}\| + \|r_T\|,$$

where $\varepsilon_{-k} = [\varepsilon_{1-k}, ..., \varepsilon_{T-k}]$ and $r_T = \sum_{k=T+1}^{\infty} \Pi_k \varepsilon_{-k}$. Obviously, for any $k = 0, ..., T$, $\|\varepsilon_{-k}\| \leq \|\varepsilon_+\|$, where $\varepsilon_+ = [\varepsilon_{1-T}, ..., \varepsilon_T]$. Latala's (2004, Thm. 2) inequality implies that $\|\varepsilon_+\| = O_{\mathrm{P}}\left(T^{1/2} + N_\varepsilon^{1/2}\right)$. Therefore,

$$\|Z\| \leq O_{\mathrm{P}}\left(T^{1/2} + N_\varepsilon^{1/2}\right) \sum_{k=0}^{T} \|\Pi_k\| + \|r_T\| = O_{\mathrm{P}}\left(T^{1/2} N^\alpha + N_\varepsilon^{1/2} N^\alpha\right) + \|r_T\|. \quad (32)$$

On the other hand,

$$\begin{aligned} \mathbb{E}\|r_T\|^2 &\leq \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[(r_T)_{it}^2\right] = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{k=T+1}^{\infty} \sum_{s=1}^{N_\varepsilon} (\Pi_k)_{is} \varepsilon_{s,t-k}\right]^2 \\ &\leq T \sum_{k=T+1}^{\infty} \|\Pi_k\|_F^2 = O\left(N_\varepsilon N^{2\alpha}\right). \end{aligned}$$

Hence, $\|r_T\| = O_{\mathrm{P}}\left(N_\varepsilon^{1/2} N^\alpha\right)$. Combining this with (32) yields (31). □

Remark 8 *The lemma holds under the following simpler, but stronger, assumptions.*

$$\sum_{k=0}^{\infty} \|\Pi_k\| = O\left(N^\alpha\right) \text{ and } \sum_{k=0}^{\infty} k \|\Pi_k\|^2 = O\left(N^{2\alpha}\right).$$

This follows from the inequalities $\|\Pi_k\|_F^2 \leq \min\{N, N_\varepsilon\} \|\Pi_k\|^2$ *and* $T \sum_{k=T+1}^{\infty} \|\Pi_k\|^2 \leq \sum_{k=0}^{\infty} k \|\Pi_k\|^2$.

By definition of Ψ_k^*, we have

$$\sum_{k=0}^{\infty} \|\Psi_k^*\| \leq \sum_{k=0}^{\infty} k \|\Psi_k\| = O\left(N^\alpha\right),$$

where the latter equality holds by A2. Further,

$$k \|\Psi_k^*\| \leq k \sum_{j=k+1}^{\infty} \|\Psi_j\| \leq \sum_{j=k+1}^{\infty} j \|\Psi_j\| = O\left(N^\alpha\right).$$

Therefore, $\sum_{k=0}^{\infty} k \|\Psi_k^*\|^2 = O\left(N^{2\alpha}\right)$. Hence, by Remark 8,

$$\|\Psi^*(L)\varepsilon M\|^2 \leq \|\Psi^*(L)\varepsilon\|^2 = O_{\mathrm{P}}\left(TN^{2\alpha} + N_\varepsilon N^{2\alpha}\right). \quad (33)$$

By assumption of Theorem 1, the right hand side of (33) is dominated by $T^2 \operatorname{tr} W = T^2 \operatorname{tr} \Omega$, which implies that statements (i) and (ii) of the theorem remain valid when $\tilde{\lambda}_k$ and $\tilde{F}_k$ are replaced by $\hat{\lambda}_k$ and $\hat{F}_k$.

To show that (iii) holds for $\hat{\lambda}_k$ and $\hat{\Sigma}$ if it holds for $\tilde{\lambda}_k$ and $\tilde{\Sigma}$, we need to establish asymptotic equivalence of $\operatorname{tr}\hat{\Sigma} = \sum_{i=1}^{T} \hat{\lambda}_i$ and $\operatorname{tr}\tilde{\Sigma} = \sum_{i=1}^{T} \tilde{\lambda}_i$. From (11),

$$\left|\hat{\lambda}_i^{1/2} - \tilde{\lambda}_i^{1/2}\right| \leq \left\|\Psi^*(L)\,\varepsilon M\right\| / \sqrt{N} \text{ and } \hat{\lambda}_i = \tilde{\lambda}_i = 0 \text{ for } i > \min\{N, T\}.$$

Therefore, by Minkowski's inequality,

$$\left|\left(\operatorname{tr}\hat{\Sigma}\right)^{1/2} - \left(\operatorname{tr}\tilde{\Sigma}\right)^{1/2}\right| \leq \left\|\Psi^*(L)\,\varepsilon M\right\| \min\left\{1, \sqrt{T/N}\right\}, \tag{34}$$

and

$$\begin{aligned}\left|\operatorname{tr}\hat{\Sigma} - \operatorname{tr}\tilde{\Sigma}\right| \;\leq\; & 2\left\|\Psi^*(L)\,\varepsilon M\right\| \min\left\{1, \sqrt{T/N}\right\}\left(\operatorname{tr}\tilde{\Sigma}\right)^{1/2} \\ & + \left\|\Psi^*(L)\,\varepsilon M\right\|^2 \min\{1, T/N\}.\end{aligned}$$

Using (27) and (33), we conclude that

$$\begin{aligned}\left|\operatorname{tr}\hat{\Sigma} - \operatorname{tr}\tilde{\Sigma}\right| \;\leq\; & T\min\left\{1, \sqrt{T/N}\right\} O_{\mathrm{P}}\left(T^{1/2}N^{\alpha} + N_{\varepsilon}^{1/2}N^{\alpha}\right)\left(\operatorname{tr}W/N\right)^{1/2} \\ & + O_{\mathrm{P}}\left(TN^{2\alpha} + N_{\varepsilon}N^{2\alpha}\right)\min\{1, T/N\}.\end{aligned}$$

It remains to show that, under the assumption made in (iii), the right hand side of the latter equality is asymptotically dominated by $\operatorname{tr}\tilde{\Sigma}$. By (27), such an asymptotic domination takes place if

$$\left(\operatorname{tr}W\right)^{-1} = o\left(\frac{T^2}{\min\{N, T\}\left(T + N_{\varepsilon}\right)N^{2\alpha}}\right).$$

But this is equivalent to the assumption made in (iii) because $\operatorname{tr}W = \operatorname{tr}\Omega$.

References

[1] Bai, J. (2004) "Estimating cross-section common stochastic trends in nonstationary panel data," *Journal of Econometrics* 122, 137–183.

[2] Bai, J. and S. Ng (2004) "A PANIC attack on unit roots and cointegration," *Econometrica* 72, 1127–1177.

[3] Banerjee, A., Marcellino, M., and Masten, I. (2017) "Structural FECM: Cointegration in large-scale structural FAVAR models," *Journal of Applied Econometrics* 32, 1069–1086.

[4] Barigozzi, M., M. Lippi, and M. Luciani (2018) "Non-Stationary Dynamic Factor Models for Large Datasets," *arXiv 1602.02398v3*

[5] Boivin, J., Giannoni, M. P., and Mojon, B. (2009) "How Has the Euro Changed the Monetary Transmission Mechanism?" in Acemoglu, D., Rogoff, K., and Woodford, M. (eds.) *NBER Macroeconomic Annual 2008*, Volume 23.

[6] Böttcher, A. and Silbermann, B. (1999) *Introduction to Large Truncated Toeplitz Matrices*, Springer.

[7] Bryzgalova, S. (2018) "Spurious Factors in Linear Asset Pricing Models," *working paper*, Stanford Graduate School of Business.

[8] Corielli, F. and Marcellino, M. (2006) "Factor based index tracking," *Journal of Banking & Finance* 30, 2215–2233.

[9] Eickmeier, S. (2009) "Comovements and Heterogeneity in the Euro Area Analyzed in a Non-stationary Dynamic Factor Model," *Journal of Applied Econometrics* 24, 933–959.

[10] Engel, C., N.C. Mark, and K.D. West (2015) "Factor Model Forecasts of Exchange Rates," *Econometric Reviews* 34, 32-55.

[11] Ghate, C. and S. Wright (2012) "The "V-factor": Distribution, timing and correlates of the great Indian growth turnaround," *Journal of Development Economics* 99, 58–67.

[12] Kato, T. (1980) *Perturbation Theory for Linear Operators*, Springer-Verlag.

[13] Koltchinskii, V. and Lounichi, K. (2016) "Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance," *Annales de l'Institut Henri Poincaré* 52, 1976–2013.

[14] Latala, R. (2004) "Some Estimates of Norms of Random Matrices," *Proceedings of the American Mathematical Society* 133, 1273–1282.

[15] McCracken, M. W. and Ng, S. (2015) "FRED-MD: A monthly database for macroeconomic research," *Working Papers 2015-12*, St. Louis, MO, Federal Reserve Bank of St. Louis.

[16] Moon, H. R. and Perron, B. (2007) "An Empirical Analysis of Nonstationarity in a Panel of Interest Rates with Factors," *Journal of Applied Econometrics* 22, 383–400.

[17] Müller, U. K. and Watson, M. W. (2008) "Testing Models of Low-Frequency Variability," *Econometrica* 76, 979–1016.

[18] Onatski, A. (2015) "Asymptotic Analysis of the Squared Estimation Error in Misspecified Factor Models," *Journal of Econometrics* 186, 388–406.

[19] Onatski, A. and Wang, C. (2020) "Spurious Factors in Data with Local-to-unit Roots," in preparation.

[20] Phillips, P.C.B. (1986) "Understanding Spurious Regression in Econometrics," *Journal of Econometrics* 33, 311–340.

[21] Phillips, P.C.B. (1998) "New tools for understanding spurious regression," *Econometrica* 66, 1299-1325.

[22] Shorack, G. R., and Wellner, J. A. (1986) *Empirical Processes with Applications to Statistics.* New York: Wiley.

[23] Stock, J.H., and Watson, M.W. (2016) "Factor Models and Structural Vector Autoregressions in Macroeconomics," in *Handbook of Macroeconomics*, Vol2A, John B. Taylor and Harald Uhlig (eds), Chapter 8, 415–526.

[24] Uhlig, H. (2009) "Comment on 'How Has the Euro Changed the Monetary Transmission Mechanism?'," in Acemoglu, D., Rogoff, K., and Woodford, M. (eds) *NBER Macroeconomic Annual 2008*, Volume 23.

[25] Vershynin, R. (2012) "Introduction to the non-asymptotic analysis of random matrices," in Eldar, Y. and Kutyniok, G. *Compressed Sensing, Theory and Applications.* Cambridge University Press, 210–268.

[26] von Borstel, J., Eickmeier, S., and Krippner, L. (2016) "The interest rate pass-through in the euro area during the sovereign debt crisis," *Journal of International Money and Finance* 68, 386–402.

[27] Wang, Y. C. and Wu J. L. (2015) "Fundamentals and Exchange Rate Prediction Revisited," *Journal of Money, Credit and Banking* 47, 1651–1671.

[28] West, K. D. and Wong, K. F. (2014) "A factor model for co-movements of commodity prices," *Journal of International Money and Finance* 42, 289–309.