ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

An algebraic estimator for large spectral density matrices

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
An algebraic estimator for large spectral density matrices / Barigozzi, Matteo; Farne, Matteo. - In: JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION. - ISSN 0162-1459. - ELETTRONICO. - NA:(2022), pp. 1-37. [10.1080/01621459.2022.2126780]

Availability:
This version is available at: https://hdl.handle.net/11585/894403 since: 2023-07-06

Published:
DOI: http://doi.org/10.1080/01621459.2022.2126780

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

25 April 2024

This is the final peer-reviewed accepted manuscript of:

Barigozzi, M., & Farnè, M. (2022). An algebraic estimator for large spectral density matrices. *Journal of the American Statistical Association*, 1-13.

The final published version is available online at:

https://doi.org/10.1080/01621459.2022.2126780

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

An algebraic estimator for large spectral density matrices

Matteo Barigozzi
Department of Economics, University of Bologna
and
Matteo Farnè*
Department of Statistics, University of Bologna

September 16, 2022

Abstract

We propose a new estimator of high-dimensional spectral density matrices, called ALgebraic Spectral Estimator (ALSE), under the assumption of an underlying low rank plus sparse structure, as typically assumed in dynamic factor models. The ALSE is computed by minimizing a quadratic loss under a nuclear norm plus l_1 norm constraint to control the latent rank and the residual sparsity pattern. The loss function requires as input the classical smoothed periodogram estimator and two threshold parameters, the choice of which is thoroughly discussed. We prove consistency of ALSE as both the dimension p and the sample size T diverge to infinity, as well as the recovery of latent rank and residual sparsity pattern with probability one. We then propose the UNshrunk ALgebraic Spectral Estimator (UNALSE), which is designed to minimize the Frobenius loss with respect to the pre-estimator while retaining the optimality of the ALSE. When applying UNALSE to a standard US quarterly macroeconomic dataset, we find evidence of two main sources of comovements: a real factor driving the economy at business cycle frequencies, and a nominal factor driving the higher frequency dynamics. The paper is also complemented by an extensive simulation exercise.

Keywords: Large spectral density matrix, Generalized Dynamic Factor Model, Dynamic rank, Sparsity.

*Corresponding author: `matteo.farne@unibo.it`.

1 Introduction

An appealing, natural, and classical way to model time series data is through spectral analysis (Brillinger, 2001). Given a p-dimensional vector stochastic process, its $p \times p$ spectral density matrix characterizes all second order dependencies. Moreover, conditional second order dependencies can also be extracted starting from the inverse of the spectral density matrix. The spectral approach is appealing since, once we move from the time domain to the frequency domain, data become asymptotically independent, as the sample size T grows to infinity.

Statistical methods for the study of time series based on spectral analysis include: pseudo-maximum likelihood estimation (Velasco and Robinson, 2000), linear regression (Harvey, 1978), cointegration tests or information criteria based on the zero-frequency spectral density matrix of a vector of time series (Stock and Watson, 1988), and similarly seasonal cointegration tests based on the spectral density matrix at selected frequencies (Joyeux, 1992), de-trending methods (Corbae et al., 2002), Granger causality tests (Breitung and Candelon, 2006; Farnè and Montanari, 2021), and the analysis of low frequency co-movements (Müller and Watson, 2018). Finally, the inverse spectral density matrix is at the basis of graphical models and dynamic network analysis (Dahlhaus, 2000; Eichler, 2007; Davis et al., 2016).

The use of spectral analysis is widespread in many applied fields. Examples are the construction of business cycle indicators in macroeconomics (Altissimo et al., 2010), portfolio optimization at different horizons in finance (Chaudhuri and Lo, 2015), and the study of brain activity in biostatistics (Ombao et al., 2001, 2005; Fiecas and Ombao, 2011, 2016).

All above methods and applications require as input an estimator of the spectral density matrix or of its inverse. Just like for the covariance matrix estimation in time domain, estimation of a spectral density matrix is a hard problem when the dimension of the process p is comparable, or even larger, than the sample size T. In this case, the classical smoothed periodogram estimator is not positive definite simply due to lack of degrees of freedom. Given the increased availability of large datasets in the recent years, this issue becomes of fundamental importance. Wu and Zaffaroni (2018) provide consistency results for the smoothed periodogram estimator in high dimension, which hold uniformly over all frequencies. Zhang and Wu (2021) improve on those results, by providing bounds for the estimation error which hold uniformly over all p^2 entries.

To solve the problem of the curse of dimensionality, here, we start from observing that the second moments of most high-dimensional time series tend to have both a low rank and a sparse component. Indeed, on the one hand, most economic datasets are known to be mainly "dense" rather than sparse (Giannone et al., 2021). Moreover, there exist mathematical results proving that large dimensional panels of time series can in general be represented as having a factor structure (Forni and Lippi, 2001; Hallin and Lippi, 2013). On the other hand, once the common factors are controlled for, there is evidence of sparseness in the second order structure of the residuals (Barigozzi and Hallin, 2017).

In this paper, we assume that the spectral density matrix, $\Sigma(\theta)$, of p-dimensional time series has the low rank plus sparse structure: $\Sigma(\theta) = L(\theta) + S(\theta)$, $\theta \in [-\pi, \pi]$, where $L(\theta)$ has rank $r < p$, and $S(\theta)$ is a sparse matrix. Based on this assumption, our estimators $\widehat{L}(\theta)$ and $\widehat{S}(\theta)$ of the two components of the spectral density matrix are obtained by regularizing the smoothed periodogram estimator, $\widetilde{\Sigma}(\theta)$, by means of a nuclear norm plus l_1 norm penalization. Specifically, at each given frequency $\theta \in [-\pi, \pi]$ we define the ALgebraic Spectral Estimator (ALSE) as the couple of matrices

$$\left(\widehat{L}(\theta), \widehat{S}(\theta)\right) = \underset{\underline{L}(\theta), \underline{S}(\theta)}{\operatorname{argmin}} \frac{1}{2} \|\widetilde{\Sigma}(\theta) - (\underline{L}(\theta) + \underline{S}(\theta))\|_F^2 + \psi \|\underline{L}(\theta)\|_* + \rho \|\underline{S}(\theta)\|_1, \tag{1}$$

where $\underline{L}(\theta)$ and $\underline{S}(\theta)$ indicate generic values of the matrices belonging to appropriate algebraic matrix varieties, $\|\underline{L}(\theta)\|_* = \operatorname{tr}(\underline{L}(\theta))$ and $\|\underline{S}(\theta)\|_1 = \sum_{i,j=1}^{p} |\underline{S}_{ij}(\theta)|$, and ψ and ρ are threshold parameters. An estimator of the spectral density matrix is then $\widehat{\Sigma}(\theta) = \widehat{L}(\theta) + \widehat{S}(\theta)$. The above optimization problem is solved by iterating between a singular value thresholding step (Cai et al., 2010), giving $\widehat{L}(\theta)$, and a soft-thresholding step (Daubechies et al., 2004), giving $\widehat{S}(\theta)$. The algorithm we employ has also been described in Luo (2011) for the case of covariance estimation.

We prove the algebraic and parametric consistency of ALSE uniformly over frequencies, as both the dimension p and the sample size T diverge. By algebraic consistency, we mean that, with probability tending to 1: (i) the ALSE low rank estimate is positive semidefinite having as rank the true rank r, (ii) the ALSE residual estimate is positive definite having as sparsity pattern the true one, and (iii) the ALSE estimate of $\Sigma(\theta)$ is positive definite. Our consistency results are obtained by generalizing to our framework the results for the smoothed periodogram of Wu and Zaffaroni (2018) and Zhang and Wu (2021). We also provide a thorough discussion on the selection of the threshold parameters ψ and ρ.

We then apply to ALSE in (1) the un-shrinkage step of estimated latent eigenvalues by Farnè and Montanari (2020), which optimizes the finite sample Frobenius loss with respect to the smoothed periodogram while retaining algebraic consistency. This step consists in unshrinking $\widehat{L}(\theta)$, while a new estimate of the sparse component is obtained by retaining the off-diagonal sparsity pattern of $\widehat{S}(\theta)$, and computing its diagonal by difference from the diagonal of $\widehat{\Sigma}(\theta)$ and the new estimate of the low rank component. We call the resulting estimator of the overall spectral density matrix UNshrunk ALgebraic Spectral Estimator (UNALSE). By construction UNALSE improves over ALSE in terms of Frobenius norm while it is equivalent to ALSE in terms of spectral norm.

Our approach is based on the fundamental identifiability assumptions we make on the behavior of the eigenvalues of the spectral density matrix. We assume the r eigenvalues of the low rank component, $L(\theta)$, to be diverging at a rate p^α with $\alpha \in [0, 1]$. In the language of factor models, this means we are allowing for the presence of factors with different degrees of pervasiveness, i.e., both weak and strong factors. Moreover, we assume the sparse component, $S(\theta)$, to have eigenvalues diverging at most at a rate p^δ with $\delta \in [0, 1/2]$ and $\delta < \alpha$. These assumptions imply the existence of an eigen-gap in the spectrum of the spectral density matrix, $\Sigma(\theta)$, which widens as p increases.

There exist alternative approaches to the estimation of large spectral density matrices. Forni et al. (2000) propose principal component analysis in the frequency domain to recover the low rank component. Böhm and von Sachs (2008, 2009) propose to shrink the smoothed periodogram towards either a reduced rank target or the identity, respectively. Fiecas and von Sachs (2014) propose a penalized likelihood approach, and Fiecas et al. (2019) consider constrained l_1 minimization for estimating the inverse. While some of those works assume either a low rank or a sparsity structure, none of them considers both assumptions jointly.

Similar approaches based on a low rank plus sparse assumption exist in time domain, i.e., for the estimation of the covariance matrix. Fan et al. (2013) consider principal components to recover the low rank component and then, in a second step, apply soft or hard thresholding to the orthogonal complement to obtain a sparse and positive definite residual. Their resulting estimator is called POET.

Farnè and Montanari (2020) adopt a minimization algorithm analogous to the one considered in this paper which recovers the covariance matrix consistently, both algebraically and parametrically. The resulting estimator is called UNALCE and they show that it systematically outperforms POET both in terms of parametric consistency, and, more importantly, because it provides the algebraic recovery of latent rank and sparsity pattern. A similar approach was proposed by Luo (2011), however it is based on the assumption of bounded eigenvalues for the covariance matrix, which does not allow for the joint identification of the two components.

Our assumption of a low rank plus sparse decomposition of the spectral density matrix is also strictly related to, and inspired by, the Generalized Dynamic Factor Model (GDFM) representation of a large panel of time series, originally proved by Forni and Lippi (2001). This is a very popular approach to dimension reduction (see, e.g., the application in Altissimo et al., 2010). In the GDFM, r latent factors are loaded by each series in a dynamic way, i.e., not only contemporaneously but also with lags. The key assumptions are: (i) pervasiveness of the factors resulting in r leading spiking spectral eigenvalues, and (ii) weak serial and cross-correlation in the residuals, resulting in boundedness of the spectral eigenvalues. In our notation these conditions imply $\alpha = 1$ and $\delta = 0$.

Forni et al. (2000, 2005, 2017) consider different estimators of the GDFM, which are all built starting from a consistent estimator of the spectral density matrix. In particular, in all those approaches the low rank component of the spectral density is estimated via the r leading dynamic principal components, i.e., the principal components of the spectral density matrix across frequencies of the smoothed periodogram (see also Brillinger, 2001). The consistency of this method relies on the pervasiveness of spectral eigenvalues with respect to the dimension p. The spectral density of the residual component, called idiosyncratic component in the GDFM literature, is then estimated as the difference between the estimated spectral density of the observed data and its estimated low rank component. Hence, by construction, the spectral density of the idiosyncratic component has rank $p - r$, i.e., it is not positive definite, and, therefore, not invertible. There exist also few papers dealing with determining the dynamic rank, r: Hallin and Liška (2007) propose an information criterion, and Onatski (2009) proposes a test based on the asymptotic distribution of the spectral eigenvalues.

The above approaches to the estimation of the GDFM suffer from some drawbacks. First, any estimator of the spectral density matrix based on the principal components of an input estimator, like the smoothed periodogram, is likely to suffer from numerical instability, especially if p is large, due to the Marčenko and Pastur (1967) law. Second, the strict pervasiveness assumption of spectral eigenvalues ($\alpha = 1$) is rarely satisfied in practice, since the factor strength might vary across frequencies, e.g., due to common, frequency specific, features. Third, the weak correlation assumption increases the number of parameters when p is large, which prevents the residual component to be identified.

The estimator we propose in this paper is able to address those drawbacks, because it tolerates weakly pervasive factors, relevantly reduces the number of estimated parameters and, given its algebraic consistency, it is also a consistent estimator of the latent rank r. For these reasons, it can be used as input of all the estimators of the GDFM considered in the literature.

The paper is organized as follows. In Section 2 we present our main results using the GDFM setting as a guiding example. In Sections 3, 4, and 5 we present the general framework, describe estimation, and prove consistency. In Section 6 we discuss the unshrinking of ALSE generating UNALSE. In Section 7 we discuss the choice of the threshold parameters. Section 8 presents the results when applying UNALSE to a dataset of quarterly US macroeconomic time series. In Appendix A we prove all theoretical results of the paper. In Appendix B we consider the implications of our assumptions for a large VAR setting. In Appendix C we show how the theory presented can be extended also to non-linear models. In Appendix D we show simulation results under a variety of data generating processes.

Notation. We denote a $p \times p$ Hermitian positive-definite complex matrix as A and its transposed complex conjugate as $A^\dagger$. Let $\lambda_i(A)$, $i = 1, \ldots, p$, be the (real) eigenvalues of A in descending order, and by A_{ij} its (i, j)th entry. Define $\overline{A}_{ij}$ as the complex conjugate of A_{ij}, and $|A_{ij}| = \sqrt{A_{ij}\overline{A}_{ij}}$; the real and imaginary parts are indicated as $\mathrm{Re}(A_{ij})$ and $\mathrm{Im}(A_{ij})$, respectively. To indicate that A is positive definite or semidefinite we write: $A \succ 0$ or $A \succeq 0$, respectively. We use the following norms. Element-wise norms: l_0 norm: $\|A\|_0 = \sum_{i=1}^p \sum_{j=1}^p \mathbb{1}(A_{ij} \neq 0)$; l_1 norm: $\|A\|_1 = \sum_{i=1}^p \sum_{j=1}^p |A_{ij}|$; Frobenius norm: $\|A\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^p |A_{ij}|^2} = \sqrt{\mathrm{tr}(AA^\dagger)}$; maximum norm: $\|A\|_\infty = \max_{1 \leq i,j \leq p} |A_{ij}|$. Vector-induced norms: $l_{0,v}$ norm $\|A\|_{0,v} = \max_{1 \leq i \leq p} \sum_{j=1}^p \mathbb{1}(A_{ij} \neq 0)$; $l_{1,v}$ norm $\|A\|_{1,v} = \max_{1 \leq j \leq p} \sum_{i=1}^p |A_{ij}|$; $l_{\infty,v}$ norm $\|A\|_{\infty,v} = \max_{1 \leq i \leq p} \sum_{j=1}^p |A_{ij}|$; spectral norm: $\|A\|_2 = \sqrt{\lambda_1(AA^\dagger)} = \lambda_1(A)$; nuclear norm: $\|A\|_* = \mathrm{tr}(A) = \sum_{i=1}^p \lambda_i(A)$. The minimum nonzero off-diagonal element of A in absolute value is denoted as $\|A\|_{min,off} = \min_{\substack{1\le i,j\le p\\ i\neq j, A_{ij}\neq 0}} |A_{ij}|$.

For two sequences $\{a_n\}$ and $\{b_n\}$ such that $a_n, b_n \to \infty$ as $n \to \infty$, we write $a_n = O(b_n)$ if $a_n/b_n \le C$ for some finite positive real C independent of n and $a_n \asymp b_n$ if $a_n/b_n = O(1)$ and $b_n/a_n = O(1)$.

2 Overview of results

In this section, we present the main features of our estimator under the assumption that the data follow a Generalized Dynamic Factor Model (GDFM) as defined by Forni and Lippi (2001). The GDFM setting has to be considered just as a motivating example, which is well suited to allow the reader to immediately appreciate the contribution of this paper with respect to the state of art. In the following sections, we present our theory in more detail showing that the validity of our results is actually much broader than the case here considered.

Let $X = \{X_{it}, i = 1, \dots, p\,, t \in \mathbb{Z}\}$ be a p-dimensional panel of zero-mean second order stationary time series. The set of all L_2-convergent linear combinations of X_{it}'s and their limits, as $p \to \infty$, of L_2-convergent sequences thereof, is a Hilbert space, denoted by $\mathcal{H}_X$. Hence, for all $t \in \mathbb{Z}$ and all $p \in \mathbb{N}$, any dynamic linear combination of X_{it}s, $y_t = \sum_{i=1}^p \sum_{k=-\infty}^{\infty} a_{ik} X_{i,t-k}$, such that $\sum_{i=1}^p \sum_{k=-\infty}^{\infty} a_{ik}^2 = 1$, belongs to $\mathcal{H}_X$. Following (Hallin and Lippi, 2013, Definitions 2.1, 2.2), we define as *common variable* the L_2-limit of any standardized dynamic linear combination of the Xs, say $\frac{y_t}{\sqrt{\mathrm{Var}(y_t)}}$, such that $\mathrm{Var}(y_t) \to \infty$, as $p \to \infty$. The Hilbert space of all *common variables* is denoted by $\mathcal{H}_{com}$, while its orthogonal complement with respect to $\mathcal{H}_X$, denoted as $\mathcal{H}_{idio}$, contains all the *idiosyncratic variables*, i.e., all dynamic linear combinations y_t with bounded variance $\mathrm{Var}(y_t)$ for all $p \in \mathbb{N}$.

Hallin and Lippi (2013) prove that there exist two unique zero-mean stochastic processes $\{\chi_{it}\} \in \mathcal{H}_{com}$ and $\{\epsilon_{it}\} \in \mathcal{H}_{idio}$, mutually orthogonal at all leads and lags, such that

$$X_{it} = \chi_{it} + \epsilon_{it}, \qquad i \in \mathbb{N}, \quad t \in \mathbb{Z}. \tag{2}$$

The process $\{\chi_{it}\}$ is called *common component*, the process $\{\epsilon_{it}\}$ is called *idiosyncratic component*. We denote the $p \times p$ spectral density matrices of the p-dimensional processes $\{X_t = (X_{1t} \cdots X_{pt})'\}$,

$\{\chi_t = (\chi_{1t} \cdots \chi_{pt})'\}$, and $\{\epsilon_t = (\epsilon_{1t} \cdots \epsilon_{pt})'\}$ as $\Sigma(\theta)$, $L(\theta)$ and $S(\theta)$, respectively. Under representation (2): (i) $\{\chi_t\}$ is driven by an r-tuple of mutually orthogonal white noises loaded by a linear time filter for all $p \in \mathbb{N}$, i.e., $\mathrm{rk}(L(\theta)) = r$ for all $\theta \in [-\pi, \pi]$; (ii) $\{\epsilon_t\}$ is orthogonal to those white noises at all leads and lags, and (iii) $\{X_t\}$ follows representation (2) if and only if the r eigenvalues of $L(\theta)$ diverge for all $\theta \in [-\pi, \pi]$ as p diverges, while the eigenvalues of $S(\theta)$ remain bounded for all $p \in \mathbb{N}$ and all $\theta \in [-\pi, \pi]$ (Forni et al., 2000). The scalar r is called the number of *dynamic factors*. All this defines the GDFM, which encompasses the *approximate static* factor models of Chamberlain and Rothschild (1983), as well as the *exact dynamic* factor models of Sargent and Sims (1977).

As usual in the GDFM literature, in this section we adopt the assumption (relaxed later on) that the r eigenvalues of $\frac{L(\theta)}{p}$ are bounded away from 0 for all $p \in \mathbb{N}$ and all $\theta \in [-\pi, \pi]$. Similarly, the definition of idiosyncratic variable leads to the condition $\lambda_1(S(\theta)) = \|S(\theta)\|_2 < \infty$ for all $p \in \mathbb{N}$ and all $\theta \in [-\pi, \pi]$. These assumptions on $L(\theta)$ and $S(\theta)$ imply that the gap between the rth and the $(r+1)$th eigenvalue of the spectral density matrix $\Sigma(\theta)$ increases at all $\theta \in [-\pi, \pi]$ as p increases, making the recovery of the low rank component possible in the limit $p \to \infty$.

In this section, we further adopt the assumption (relaxed later on) that the idiosyncratic spectral density matrix $S(\theta)$ is such that $\|S(\theta)\|_{0,v}$ is bounded for all $p \in \mathbb{N}$ and all $\theta \in [-\pi, \pi]$. Since $\|S(\theta)\|_2 \leq \|S(\theta)\|_{0,v}$, the original assumption $\|S(\theta)\|_2 < \infty$ still holds. This is done in order to enforce element-wise sparsity on $S(\theta)$.

Suppose now that we observe a sample of size T of p-dimensional data vectors. A classical estimator of the spectral density matrix, which is our *pre-estimator*, is the smoothed periodogram, defined as

$$\widetilde{\Sigma}(\theta_h) = \frac{1}{2\pi} \sum_{k=-(T-1)}^{T-1} K\left(\frac{k}{M_T}\right) \mathrm{e}^{-\mathrm{i}\theta_h k}\, \widehat{\Gamma}_X(k), \qquad \theta_h = \frac{h\pi}{M_T}, \quad |h| \leq M_T, \tag{3}$$

where $\widehat{\Gamma}_X(k) = T^{-1} \sum_{t=1}^{T-|k|} X_t X'_{t+k}$, and $K(\cdot)$ is a suitable kernel function with M_T being the associated smoothing parameter. According to Brillinger (2001), for any given θ_h, $\widetilde{\Sigma}(\theta_h)$ is consistent if $\frac{M_T}{T} \to 0$ while $M_T \to \infty$ and $T \to \infty$. Wu and Zaffaroni (2018) prove the consistency of $\widetilde{\Sigma}(\theta_h)$ uniformly over the frequencies, under appropriate assumptions to be discussed later.

Under the GDFM setting described above, augmented with the sparsity assumption for $S(\theta)$, we define the ALgebraic Spectral Estimator (ALSE) estimator of the spectral density matrix $\Sigma(\theta)$ as $\widehat{\Sigma}(\theta) = \widehat{L}(\theta) + \widehat{S}(\theta)$, where $\widehat{L}(\theta)$ and $\widehat{S}(\theta)$ are such that they satisfy (1). We refer to Section 4 for details on computing the solution of such minimization problem.

Consistency of ALSE under the GDFM setting is in the following Corollary to our main contribution which is Theorem 5.1.

Corollary 2.1. *For all $p \in \mathbb{N}$, assume that: (i) the r nonzero eigenvalues of $L(\theta)$ are such that $\frac{\lambda_j(L(\theta))}{p}$ is finite and bounded away from zero for all $j = 1, \dots, r$; (ii) $\|S(\theta)\|_{0,v}$ is bounded; (iii) $\{\chi_{it}, t \in \mathbb{Z}, i = 1, \dots, p\}$ and $\{\epsilon_{it}, t \in \mathbb{Z}, i = 1, \dots, p\}$, are Gaussian processes. Then, under the assumptions and the conditions of Theorem 5.1, there exist positive reals κ_1 and κ_2 independent of p and T such that, as $T \to \infty$ and for all $p \in \mathbb{N}$, the following hold:*

1. $\mathcal{P}\left(\max_{|h| \le M_T} \frac{1}{p} \|\widehat{L}(\theta_h) - L(\theta_h)\|_2 \le \kappa_1 \sqrt{\frac{M_T \log(M_T p)}{T}}\right) \to 1;$
2. $\mathcal{P}\left(\mathrm{rk}(\widehat{L}(\theta_h)) = r\right) \to 1$, *for all* $\theta_h = \frac{h\pi}{M_T}$, $|h| \le M_T;$
3. $\mathcal{P}\left(\max_{|h| \le M_T} \|\widehat{S}(\theta_h) - S(\theta_h)\|_2 \le \kappa_2 \left\{\max_{|h| \le M_T} \|S(\theta_h)\|_{0,v}\right\} \sqrt{\frac{M_T \log(M_T p)}{T}}\right) \to 1;$
4. $\mathcal{P}\left(\max_{|h| \le M_T} \frac{1}{p} \|\widehat{\Sigma}(\theta_h) - \Sigma(\theta_h)\|_2 \le \kappa_1 \sqrt{\frac{M_T \log(M_T p)}{T}}\right) \to 1.$

Furthermore, under the assumptions and the conditions of Corollary 5.1, there exists a positive real κ_4 independent of p and T such that, as $T \to \infty$ and for all $p \in \mathbb{N}$, the following hold: $\mathcal{P}\left(\max_{|h| \le M_T} \|\widehat{S}^{-1}(\theta_h) - S^{-1}(\theta_h)\|_2 \le \kappa_4 \left\{\max_{|h| \le M_T} \|S(\theta_h)\|_{0,v}\right\} \sqrt{\frac{M_T \log(M_T p)}{T}}\right) \to 1.$

The results of Corollary 2.1 contribute to the GDFM literature in three ways. First, the exact rank recovery in part 2 allows to bypass the use of existing criteria for determining the number of factors, like those by Hallin and Liška (2007) and Onatski (2009). Second, we derive a consistency result also for the estimator of the idiosyncratic spectral density $\widehat{S}(\theta)$. Third, assuming that $S(\theta)$ is full rank, we obtain a result also for the estimators of its inverse. Our consistency rates are comparable to those derived by Fan et al. (2013) for the estimation of the covariance matrix of $\{X_t\}$ when this is generated by a static factor model with sparse idiosyncratic covariance.

In fact Theorem 5.1, which is our main result, holds beyond the standard GDFM assumptions. In particular, first, we relax the strict pervasiveness assumption on latent dynamic factors, by allowing the r eigenvalues of the matrix $\frac{L(\theta)}{p^\alpha}$, with $\alpha \le 1$, to be bounded away from 0 for all $p \in \mathbb{N}$ and all

$\theta \in [-\pi, \pi]$. Second, we allow for the maximum number of nonzeros per column in $S(\theta)$, $\|S(\theta)\|_{0,v}$, to be at most proportional to p^δ, with $\delta \in [0, \frac{1}{2}]$ and $\delta < \alpha$ for all $p \in \mathbb{N}$ and all $\theta \in [-\pi, \pi]$. This means that we allow the idiosyncratic spectrum to be quite far from the diagonal matrix. Our setting reduces to the GDFM one when $\alpha = 1$ and $\delta = 0$. Third, we propose an unshrinking procedure such that the resulting UNshrunk ALgebraic Spectral Estimator (UNALSE) optimizes the finite sample Frobenius loss with respect to the smoothed periodogram while retaining algebraic consistency.

3 Model setup

The aim of this paper is estimating the spectral density matrix of a p-dimensional process $X = \{X_{it}, i = 1, \dots, p,\ t \in \mathbb{Z}\}$, following the data generating process:

$$X_t = \chi_t + \epsilon_t, \qquad t \in \mathbb{Z}, \tag{4}$$

$$\chi_t = \sum_{s=0}^{\infty} B_s u_{t-s}, \qquad t \in \mathbb{Z}, \tag{5}$$

$$\epsilon_t = \sum_{s=0}^{\infty} C_s e_{t-s}, \qquad t \in \mathbb{Z}, \tag{6}$$

where X_t, χ_t, ϵ_t, and e_t are p-dimensional, u_t is r-dimensional, the B_s are $p \times r$, and the C_s are $p \times p$. While Forni and Lippi (2001) derive a two-sided singular MA representation for the process $\{\chi_t\}$, the existence of a one-sided representation (5) is proved by Hallin and Lippi (2013, Theorem 2.2). The MA representation (6) for the process $\{\epsilon_t\}$ is the usual Wold representation.

We make the two following assumptions on the MA processes in (5) and (6).

Assumption 3.1. *(i) $\{u_t, t \in \mathbb{Z}\}$ is a r-dimensional independent and identically distributed process with $\mathrm{E}[u_t] = 0_r$ and $\mathrm{E}[u_t u_t'] = I_r$ and with r finite and independent of p for all $p \in \mathbb{N}$; (ii) there exists $K_u > 0$ and $d_u > 4$ independent of j and t such that $\mathrm{E}[|u_{jt}|^{d_u}] \leq K_u$ for all $j = 1, \dots, r$; (iii) for all $p \in \mathbb{N}$, $\{e_t, t \in \mathbb{Z}\}$ is a p-dimensional independent and identically distributed process with $\mathrm{E}(e_t) = 0_p$ and $\mathrm{E}[e_t e_t'] = I_p$; (iv) there exists $K_e > 0$ and $d_e > 4$ independent of j and t such that $\mathrm{E}[|e_{jt}|^{d_e}] \leq K_e$ for all $j \in \mathbb{N}$; (v) $\{u_t\}$ and $\{e_t\}$ are two mutually independent processes.*

Assumption 3.2. *There exist* $M_\chi, M_\epsilon > 0$*,* $\rho_\chi, \rho_\epsilon \in [0, 1)$*,* $\alpha \in (\frac{1}{2}, 1]$ *and* $\delta' \in [0, \alpha)$*, such that, for all* $p \in \mathbb{N}$*: (i)* $\|B_s\|_{\infty,v} \leq M_\chi \rho_\chi^s$*; (ii)* $\|B_s\|_F \leq M_\chi \rho_\chi^s p^{\alpha/2}$*; (iii)* $\|C_s\|_{\infty,v} \leq M_\epsilon \rho_\epsilon^s$*; (iv)* $\|C_s\|_1 \leq M_\epsilon \rho_\epsilon^s p^{\delta'/2}$*.*

Under Assumptions 3.1 and 3.2, the processes $\{\chi_t\}$ and $\{\epsilon_t\}$ satisfying (5) and (6) are zero-mean linear and weakly stationary for any fixed p, and consequently the process $\{X_t\}$ also is. Similar assumptions are made by Forni et al. (2017) in a GDFM context and allow us to control the amount of physical dependence of $\{X_t\}$ across time (Wu and Zaffaroni, 2018). Notice that cross-sectional heteroskedasticity of both $\{\chi_t\}$ and $\{\epsilon_t\}$ is allowed for.

Some more comments on Assumption 3.2 are needed. First, the fact that in part (i) there is no dependence on p is natural since B_s has a finite number of columns. Second, part (ii) implies the largest eigenvalue of $B_s B_s'$ diverges with p^α, because $\|B_s\|_2^2 \leq \|B_s\|_F^2 \leq M_\chi^2 \rho_\chi^{2s} p^\alpha$, a requirement which is compatible with the idea of pervasive factors at all lags that we impose in Assumption 3.4 below. Third, part (iii) imposes finite column sums, i.e., the $\ell_{\infty,v}$ norm, for the coefficients C_s, meaning that for each given X_{it} the p innovations $e_{1t}, \dots, e_{pt}$ have a finite effect for any $p \in \mathbb{N}$, which is in agreement with the idiosyncratic nature of $\{\epsilon_t\}$ assumed in the GDFM literature. Fourth, to account for some stronger dependence in $\{\epsilon_t\}$, in part (iv) we allow the entire ℓ_1 norm of C_s to be diverging with p. This, together with part (iii), implies that the diverging behavior of those coefficients is implicitly due to the row sums, i.e., the $\ell_{1,v}$ norm. This is just a useful and natural way of parametrizing the model and, obviously, we could equivalently assume the viceversa or let both row and column sums diverge (compatibly with part (iv)). Fourth, the assumption $\delta' < \alpha$ ensures that the common component always dominates the idiosyncratic component when $p \to \infty$.

Now, using the Singular Value Decomposition, we can always write the MA coefficients as: $B_s = U_{L,s} \Lambda_{L,s} V_{L,s}$, $C_s = U_{S,s} \Lambda_{S,s} V_{S,s}$, $s \in \mathbb{Z}^+ \cup \{0\}$, where $U_{L,s}$ is $p \times r$ with $U_{L,s}' U_{L,s} = I_r$, $V_{L,s}$ is $r \times r$ with $V_{L,s} V_{L,s}' = V_{L,s}' V_{L,s} = I_r$, $\Lambda_{L,s}$ is $r \times r$ diagonal, real, and positive definite matrix of singular values. Similarly, $U_{S,s}$ is $p \times p$ with $U_{S,s} U_{S,s}' = U_{S,s}' U_{S,s} = I_p$, $V_{S,s}$ is $p \times p$ with $V_{S,s} V_{S,s}' = V_{S,s}' V_{S,s} = I_p$, $\Lambda_{S,s}$ is $p \times p$ diagonal, real, and positive definite of singular values. By means of the following assumption, we impose a low rank plus sparse structure on the filters.

Assumption 3.3. *For all* $s \in \mathbb{Z}^+ \cup \{0\}$*: (i)* $U_{L,s} = U_L$*; (ii)* $\|U_{S,s} U_{S,s}'\|_0 = q_s$ *with* $q_s \in \mathbb{N}$ *and* $q_s < p^2$

for all $p \in \mathbb{N}$.

In other words, in part (i) we allow the matrices $B_s'B_s$ to have a different condition number across all lags s, and in part (ii) we allow the matrices C_sC_s' to have a different sparsity pattern across all lags s. Although part (i) might seem restrictive in that it assumes that the space spanned by $U_{L,s}$ is the same at all lags s, it is in fact quite reasonable if, as usually assumed, the spectral density of $\{\chi_t\}$ has to have rank r at all frequencies (see below). Moreover, we notice that all results in the next section hold locally in a neighborhood of $L(\theta)$ (see also Chandrasekaran et al., 2012).

Under Assumption 3.3, the lag-k autocovariances of $\{\chi_t\}$ and of $\{\epsilon_t\}$ are given by $\Gamma_\chi(k) = \sum_{s=0}^{\infty} U_L \Lambda_{L,s}\Lambda_{L,s+k}U_L'$ and $\Gamma_\epsilon(k) = \sum_{s=0}^{\infty} U_{S,s}\Lambda_{S,s}\Lambda_{S,s+k}U_{S,s+k}'$, respectively, and the spectral density matrices of $\{\chi_t\}$ and of $\{\epsilon_t\}$ are given by $L(\theta) = \frac{1}{2\pi}U_L\left(\sum_{k=-\infty}^{\infty}\left(\sum_{s=0}^{\infty}\Lambda_{L,s}\Lambda_{L,s+k}\right)\mathrm{e}^{-\mathrm{i}\theta\mathrm{k}}\right)U_L'$ and $S(\theta) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\left(\sum_{s=0}^{\infty}U_{S,s}\Lambda_{S,s}\Lambda_{S,s+k}'U_{S,s+k}'\mathrm{e}^{-\mathrm{i}\theta\mathrm{k}}\right)$, respectively. Therefore, $\Gamma_\chi(k)$ is of rank r for all $k \in \mathbb{Z}$, and $L(\theta)$ is of rank r for all $\theta \in [-\pi,\pi]$. This is natural and standard requirement in GDFM literature (Forni et al., 2000). Moreover, since by Assumption 3.3(ii) $U_{S,s}$ must be sparse for any $s \in \mathbb{Z}^+ \cup \{0\}$, then $\Gamma_\epsilon(k)$ will be sparse, and $S(\theta)$ will also be sparse. In particular, letting $\mathcal{Q}_k = \{i,j = 1,\dots,p,\ \Gamma_{\epsilon,ij}(k) \neq 0\}$ with cardinality, say, q, and $\mathcal{Q}(\theta) = \{i,j=1,\dots,p,\ S_{ij}(\theta) \neq 0\}$, we have $\mathcal{Q}(\theta) \subseteq \cup_{k=0}^{\infty}\mathcal{Q}_k$ with cardinality at most q, for some $\theta \in [-\pi,\pi]$. This has nontrivial implications for the ability of our method to retrieve nonzero entries which are discussed in Remark 5.6 below.

Summing up, Assumption 3.3 characterizes the spectral density matrix of $\{X_t\}$, which is $\Sigma(\theta) = L(\theta) + S(\theta)$, as having a low rank plus sparse structure at all frequencies $\theta \in [-\pi,\pi]$.

To fully characterize the low rank property of $L(\theta)$, we make the following assumption, which strengthens Assumptions 3.2(ii) and 3.3(i).

Assumption 3.4. *For all* $p \in \mathbb{N}$, *all* $\theta \in [-\pi,\pi]$, *the spectral density* $L(\theta)$ *exists, and for all* $j = 1,\dots,r$, *there exist continuous functions* $M_j^{\min}(\theta), M_j^{\max}(\theta) : [-\pi,\pi] \to \mathbb{R}^+$ *such that* $M_j^{\min}(\theta) \leq \frac{\lambda_j(L(\theta))}{p^\alpha} \leq M_j^{\max}(\theta)$, *with* $\alpha \in (\frac{1}{2},1]$, *and* $M_j^{\min}(\theta) \geq M_{j+1}^{\max}(\theta)$, *for* $j = 1,\dots,(r-1)$.

The case $\alpha = 1$ corresponds to the classical assumption in factor model literature of strong pervasiveness of the latent factors (Forni et al., 2017). Here, by allowing the eigenvalues of $L(\theta)$ to be of order p^α, $\alpha \in (\frac{1}{2},1]$, we allow also for weaker factors.

To fully characterize the sparseness property of $S(\theta)$, we make the following assumption which strengthens Assumptions 3.2(iv) and 3.3(ii).

Assumption 3.5. *For all $p \in \mathbb{N}$ and all $\theta \in [-\pi, \pi]$, the spectral density $S(\theta)$ exists and: (i) there exist $\delta \in [0, \frac{1}{2}]$ and $\delta_2 > 0$, such that $\|S(\theta)\|_{0,v} \leq \delta_2 p^\delta$ for all $\theta \in [-\pi, \pi]$; (ii) there exists $\mathcal{M} > 0$ such that $\lambda_p(S(\theta)) > \mathcal{M}$.*

Part (i) controls the maximum number of nonzeros per row in $S(\theta)$, which is independent of θ because of Assumption 3.3(ii), and it is allowed to increase with p. In particular, the maximum number of nonzeros (across frequencies) in $S(\theta)$, which we denoted as q, is such that $q \leq p \sup_{\theta \in [-\pi,\pi]} \|S(\theta)\|_{0,v}$. Hence, because of Assumption 3.5, q can grow with p at a maximum rate $p^{3/2}$. This fully characterizes the sparsity conditions on $S(\theta)$. Moreover, since Assumption 3.2(iv) implies that $\|S(\theta)\|_1 = O(p^{\delta'})$, and since by Assumption 3.5 $\|S(\theta)\|_1 \leq p\|S(\theta)\|_{1,v} \leq p\|S(\theta)\|_{0,v}\|S(\theta)\|_\infty = O(p^{\delta+1})$, then we must have $\delta' \leq 1 + \delta$. In part (ii), by assuming positive definiteness of $S(\theta)$, we guarantee that also $\Sigma(\theta)$ is positive definite, because of Weyl's inequality.

Remark 3.1. From Assumption 3.5 it follows that the largest eigenvalue of $S(\theta)$ is at most of order p^δ, $\delta \in [0, \frac{1}{2}]$ since $\|S(\theta)\|_2 \leq \|S(\theta)\|_\infty \|S(\theta)\|_{0,v} = O(p^\delta)$. Moreover, by Assumption 3.4 the eigenvalues of $L(\theta)$ are of order p^α, $\alpha \in (\frac{1}{2}, 1]$, so that $\delta < \alpha$. Therefore, there exists an eigengap between the eigenvalues of $L(\theta)$ and those of $S(\theta)$ across all frequencies $\theta \in [-\pi, \pi]$. In the GDFM setting the presence of an eigengap growing with p is the condition which allows for the recovery of the number of factors r (Hallin and Liška, 2007 and Onatski, 2009). Notice that here we impose $\lambda_j(L(\theta)) \asymp p^\alpha$, while we just have an upper bound for $\lambda_1(S(\theta))$, so the eigengap implied by our assumptions is at least of order $p^{\alpha-\delta}$ but could be wider.

Remark 3.2. It is also worth noticing that in Assumption 3.4 we could easily allow for factors having different degree of pervasiveness, i.e., by assuming the eigenvalues at frequency θ to be of order $p^{\alpha(\theta)}$ (in that case, α in Assumption 3.2 would be the infimum over frequencies of all $\alpha(\theta)$). Likewise in Assumption 3.5 we could assume $\|S(\theta)\|_{0,v} \leq \delta_2 p^{\delta(\theta)}$. However, in order to keep the notation simple, we prefer to keep treating α and δ as constants (see also Remark 5.6 below for more details).

In order to study the properties of our estimator and guarantee algebraic consistency, we need to formalize further the low rank plus sparse structure. To this end we introduce the following algebraic matrix varieties for generic integers r and q:

$$\mathcal{L}(\mathsf{r}) = \{L \in \mathbb{C}^{p\times p} \mid L \succeq 0, L = UDU^\dagger, U \in \mathbb{C}^{p\times \mathsf{r}}, U^\dagger U = I_\mathsf{r}, D \in \mathbb{R}^{\mathsf{r}\times\mathsf{r}}\text{diagonal}, \mathrm{rk}(D) = \mathsf{r} < p\}, \tag{7}$$

$$\mathcal{S}(\mathsf{q}) = \{S \in \mathbb{C}^{p\times p} \mid S \succ 0, |\mathrm{supp}(S)| \leq \mathsf{q} < p^2\}, \tag{8}$$

where $\mathrm{supp}(S)$ is the orthogonal complement of $\ker(S) = \{v \in \mathbb{C}^p \mid Sv = 0\}$. In other words, $\mathcal{L}(\mathsf{r})$ is the variety of Hermitian matrices with at most rank r and $\mathcal{S}(\mathsf{q})$ is the variety of Hermitian matrices with at most q nonzero elements. Therefore, under our assumptions $L(\theta) \in \mathcal{L}(r)$ and $S(\theta) \in \mathcal{S}(q)$ for all $\theta \in [-\pi, \pi]$.

In order to give a rigorous definition of sparsity, we need to introduce further notation. The tangent spaces to $\mathcal{L}(r)$ and $\mathcal{S}(q)$ in $L(\theta)$ and $S(\theta)$ are respectively defined, for all $\theta \in [-\pi, \pi]$, as:

$$\mathcal{T}(L(\theta)) = \{M \in \mathbb{C}^{p\times p} \mid M = UY_1^\dagger + Y_2U^\dagger \mid Y_1, Y_2 \in \mathbb{C}^{p\times r}, U \in \mathbb{C}^{p\times r}, U^\dagger U = I_r, U^\dagger L(\theta)U \in \mathbb{C}^{r\times r}\text{diagonal}\},$$

$$\Omega(S(\theta)) = \{N \in \mathbb{C}^{p\times p} \mid \mathrm{supp}(N) \subseteq \mathrm{supp}(S(\theta))\}.$$

The following uncertainty principle holds (Chandrasekaran et al., 2011): if $L(\theta)$ is nearly sparse, $S(\theta)$ cannot be recovered, and if $S(\theta)$ is nearly low rank, $L(\theta)$ cannot be recovered. Therefore, in order to identify $L(\theta)$ and $S(\theta)$ at each $\theta \in [-\pi, \pi]$, we need to ensure that the tangent spaces to $\mathcal{L}(r)$ and $\mathcal{S}(q)$ are close to orthogonality. To this end, Chandrasekaran et al. (2011) introduce the following measures for any generic $L \in \mathcal{L}(r)$ and $S \in \mathcal{S}(q)$:

$$\xi(\mathcal{T}(L)) = \max_{\substack{M\in\mathcal{T}(L)\\ \|M\|_2\leq 1}} \|M\|_\infty, \qquad \mu(\Omega(S)) = \max_{\substack{N\in\Omega(S)\\ \|N\|_\infty\leq 1}} \|N\|_2.$$

So if $\xi(\mathcal{T}(L))$ is small then the elements of the tangent space $\mathcal{T}(L)$ are "diffuse", i.e., these elements are not too sparse; as a result L cannot be very sparse. Similarly, the quantity $\mu(\Omega(S))$ being small implies that the spectrum of any element of the tangent space $\Omega(S)$ is "diffuse", i.e., the singular values of S are not too large. Moreover, (Chandrasekaran et al., 2011, Propositions 3 and 4) show that the following relationships always hold for any generic $L \in \mathcal{L}(r)$ and $S \in \mathcal{S}(q)$:

$$\sqrt{\frac{r}{p}} \leq \xi(\mathcal{T}(L)) \leq 2, \qquad \min_{1\leq i\leq p} \sum_{j=1}^{p} \mathbb{1}(S_{ij}=0) \leq \mu(\Omega(S)) \leq \max_{1\leq i\leq p} \sum_{j=1}^{p} \mathbb{1}(S_{ij}=0) = \|S\|_{0,v}. \tag{9}$$

Moreover, a necessary condition to ensure both parametric and algebraic consistency is:

$$\xi(\mathcal{T}(L(\theta))) \cdot \mu(\Omega(S(\theta)) \leq \frac{1}{54}, \quad \theta \in [-\pi, \pi], \tag{10}$$

which guarantees that $L(\theta)$ is far from sparsity and $S(\theta)$ is far from rank-deficiency. Indeed, the smaller is the product between the dual norms $\xi(\mathcal{T}(L(\theta)))$ and $\mu(\Omega(S(\theta))$, the closer the two spaces $\mathcal{L}(r)$ and $\mathcal{S}(q)$ are to orthogonality, thus making easier to perform low rank plus sparse decomposition effectively.

To control the rank-sparsity incoherence measures we make the following assumption, which guarantees that (9) and (10) are satisfied.

Assumption 3.6. *For all $p \in \mathbb{N}$, there exist $\kappa_L, \kappa_S > 0$ with $\frac{\sqrt{r}\kappa_S}{\kappa_L} \leq \frac{1}{54}$, $\frac{\sqrt{r}}{2} \leq \kappa_L$ and $\kappa_S \leq \delta_2$, such that $\xi(\mathcal{T}(\theta)) = \frac{\sqrt{r}}{\kappa_L p^\delta}$ and $\mu(\Omega(\theta)) = \kappa_S p^\delta$, for all $\theta \in [-\pi, \pi]$, and where δ and δ_2 are defined in Assumption 3.5.*

Now, because of Assumption 3.3, $\mathcal{T}(L(\theta))$ depends only on the rank of $L^* = \Gamma_\chi(0)$ which is r, and $\Omega(S(\theta))$ depends only on the support of $S_\infty = \sum_{k=-\infty}^{\infty} \Gamma_\epsilon(k)$, which has dimension q. For this reason, hereafter, we use the shorthand notation $\mathcal{T}$, Ω, $\xi(\mathcal{T})$, and $\mu(\Omega)$.

Remark 3.3. The validity of our setup goes beyond the VMA(∞) framework discussed so far. First, it is obvious that any stable VARMA with finite lags would fit into our framework. We refer to Appendix B for specific conditions on the VARMA coefficients such that our assumptions are satisfied. Second, as long as $\{u_t\}$ and $\{e_t\}$ can be expressed as measurable functions of i.i.d. processes satisfying Assumption 3.1, it can be shown that the theory developed in this paper is still valid. We refer to Appendix C for details and the extension to the non-linear case in which $\{u_t\}$ and $\{e_t\}$ are allowed to be conditionally heteroskedastic.

4 Estimation

Suppose now to observe a sample of p-dimensional data vectors with size T, i.e., we observe $\{X_{it}, i = 1, \dots, p,\ t = 1, \dots, T\}$, and we compute the ALSE $\widehat{\Sigma}(\theta) = \widehat{L}(\theta) + \widehat{S}(\theta)$ such that it satisfies (1) which, for convenience, we rewrite here:

$$\left(\widehat{L}(\theta), \widehat{S}(\theta)\right) = \underset{\substack{\underline{L}(\theta)\succeq 0, \underline{S}(\theta)\succ 0\\ \underline{L}(\theta)+\underline{S}(\theta)\succ 0}}{\arg\min} \frac{1}{2}\|\widetilde{\Sigma}(\theta) - (\underline{L}(\theta) + \underline{S}(\theta))\|_F^2 + \psi\|\underline{L}(\theta)\|_* + \rho\|\underline{S}(\theta)\|_1, \tag{11}$$

where $\widetilde{\Sigma}(\theta)$ is the smoothed periodogram estimator defined in (3), $\underline{L}(\theta)$ and $\underline{S}(\theta)$ indicate generic values of the matrices, and ψ and ρ are positive threshold parameters and their choice is discussed in Section 7.

The minimization problem (11) is a non-smooth convex optimization problem which is the tightest convex relaxation of the following NP-hard problem:

$$\min_{\underline{L}(\theta), \underline{S}(\theta)} \frac{1}{2}\|\widetilde{\Sigma}(\theta) - (\underline{L}(\theta) + \underline{S}(\theta))\|_F^2 + \psi\, \mathrm{rk}(\underline{L}(\theta)) + \rho\|\underline{S}(\theta)\|_0, \tag{12}$$

which would be the natural target under the low rank plus sparse assumption. Indeed, we know that: (i) $\|\underline{S}(\theta)\|_1$ is the tightest convex relaxation of $\|\underline{S}(\theta)\|_0$ (Donoho, 2006); (ii) $\|\underline{L}(\theta)\|_*$ is the tightest convex relaxation of $\mathrm{rk}(\underline{L}(\theta))$ (Fazel et al., 2001).

In practice, the solution of (11) is computed as follows. For any given frequency $\theta_h = \frac{\pi h}{M_T}$, with $|h| \leq M_T$, we apply the following iterative procedure:

1. set $(L_0(\theta_h), S_0(\theta_h)) = \left(\frac{diag(\widetilde{\Sigma}(\theta))}{2}, \frac{diag(\widetilde{\Sigma}(\theta))}{2}\right)$, $\eta_0 = 1$, and initialize $Y_0(\theta_h) = L_0(\theta_h)$ and $Z_0(\theta_h) = S_0(\theta_h)$;
2. for $k \geq 1$, repeat:
   (a) compute $\frac{\partial \frac{1}{2}\|Y_{k-1}(\theta_h) + Z_{k-1}(\theta_h) - \widetilde{\Sigma}(\theta_h)\|_F^2}{\partial Y_{k-1}(\theta_h)} = \frac{\partial \frac{1}{2}\|Y_{k-1}(\theta_h) + Z_{k-1}(\theta_h) - \widetilde{\Sigma}(\theta_h)\|_F^2}{\partial Z_{k-1}(\theta_h)} = Y_{k-1}(\theta_h) + Z_{k-1}(\theta_h) - \widetilde{\Sigma}(\theta_h)$;
   (b) apply the *singular value thresholding* operator of Cai et al. (2010) $T_\psi(\cdot)$ to $\mathcal{E}_{Y,k}(\theta_h) = Y_{k-1}(\theta_h) - \frac{1}{2}(Y_{k-1}(\theta_h) + Z_{k-1}(\theta_h) - \widetilde{\Sigma}(\theta_h))$ and set $L_k(\theta_h) = T_\psi(\mathcal{E}_{Y,k}(\theta_h))$;
   (c) apply the *soft-thresholding* operator of Daubechies et al. (2004) $T_\rho(\cdot)$ to $\mathcal{E}_{Z,k}(\theta_h) = Z_{k-1}(\theta_h) - \frac{1}{2}(Y_{k-1}(\theta_h) + Z_{k-1}(\theta_h) - \widetilde{\Sigma}(\theta_h))$ and set $S_k(\theta_h) = T_\rho(\mathcal{E}_{Z,k}(\theta_h))$;
   (d) set $(Y_k(\theta_h), Z_k(\theta_h)) = (L_k(\theta_h), S_k(\theta_h)) + \frac{\eta_{k-1}-1}{\eta_k}[(L_k(\theta_h), S_k(\theta_h)) - (L_{k-1}(\theta_h), S_{k-1}(\theta_h))]$ where

$\eta_k = \frac{1+\sqrt{1+4\eta_{k-1}^2}}{2}$;

(e) stop if $\frac{\|L_k - L_{k-1}\|_F}{\|1+L_{k-1}\|_F} + \frac{\|S_k - S_{k-1}\|_F}{\|1+S_{k-1}\|_F} \leq \varsigma$, where ς is a prescribed precision level (we set $\varsigma = 0.01$);

3. set $\widehat{L}(\theta_h) = Y_k(\theta_h)$ and $\widehat{S}(\theta_h) = Z_k(\theta_h)$.

The two thresholding operators introduced in the above algorithm are defined as follows.

(I) Singular value thresholding operator: let the Singular Value Decomposition of a positive semi-definite complex symmetric matrix A be $A = U_A \Lambda_A U_A^\dagger$, then, define $T_\psi(A) = U_A \Lambda_{\psi,A} U_A^\dagger$, where $\Lambda_{\psi,A}$ is a diagonal matrix with ith diagonal element $\Lambda_{\psi,A,ii} = \max(\Lambda_{A,ii} - \psi, 0)$.

(II) Soft-thresholding operator: for a positive definite complex symmetric matrix A define $T_\rho(A_{ij}) = \frac{A_{ij}}{(A_{ij}\overline{A}_{ij})^{1/2}} \max((A_{ij}\overline{A}_{ij})^{1/2} - \rho, 0)$.

Two features of ALSE must be stressed. First, not only ALSE produces the estimates $\widehat{L}(\theta)$ and $\widehat{S}(\theta)$, but it also produces estimates of $\mathrm{rk}(L(\theta)) = r$ and of $\mathrm{supp}(S(\theta))$ and therefore of its cardinality q. Second, the solution of the above minimization can be searched without the need of constraining $\underline{L}(\theta)$ and $\underline{S}(\theta)$ to the manifolds $\mathcal{L}(r)$ and $\mathcal{S}(q)$ defined in (7)-(8) (see also Remark 5.6 and Appendix A).

5 Consistency

We now prove the algebraic and parametric consistency of the pair of estimates $(\widehat{L}(\theta), \widehat{S}(\theta))$, and, in order to do this, we introduce two definitions, taken from Chandrasekaran et al. (2012). First, we say that $(\widehat{S}(\theta), \widehat{L}(\theta))$ is algebraically consistent if the following conditions hold, for any given $\theta \in [-\pi, \pi]$: 1. $\mathrm{rk}(\widehat{L}(\theta)) = \mathrm{rk}(L(\theta))$; 2. $\mathrm{sgn}(\mathrm{Re}(\widehat{S}_{ij}(\theta))) = \mathrm{sgn}(\mathrm{Re}(S_{ij}(\theta)))$, for all $i, j = 1, \ldots, p$ (by convention we let $\mathrm{sgn}(0) = 0$); 3. $\widehat{L}(\theta) + \widehat{S}(\theta)$ and $\widehat{S}(\theta)$ are positive definite and $\widehat{L}(\theta)$ is positive semidefinite. Condition 2 is also often referred to as sparsistency (Lam and Fan, 2009). Second, we use the classical definition of parametric consistency, which holds if the estimates $(\widehat{S}(\theta), \widehat{L}(\theta))$ are close to $(S(\theta), L(\theta))$, for any given $\theta \in [-\pi, \pi]$, with high probability in ℓ_2 norm.

In order to prove consistency of the pre-estimator of the spectral density matrix $\widetilde{\Sigma}(\theta)$, defined in (3), we make the following standard assumption on the kernel function and its bandwidth.

Assumption 5.1. *The kernel function $K(\cdot)$ is even, bounded, with support $[-1, 1]$, and bandwidth M_T, such that: (i) for some $\mathsf{k} > 0$, $|K(s) - 1| = O(s^{\mathsf{k}})$, as $s \to 0$; (ii) $\int_{-\infty}^{\infty} K^2(s)ds < \mathcal{K}$ for some finite $\mathcal{K}$;*

(iii) $\sum_{s'\in\mathbb{Z}}\sup_{|s'-s''|\le 1}|K(s'\vartheta)-K(s''\vartheta)|=O(1)$*, as* $\vartheta\to 0$*; (iv)* $c_1 T^{\underline{\zeta}}\le M_T\le c_2 T^{\zeta}$*, for some* $c_1,c_2>0$ *and* $\zeta,\underline{\zeta}>0$*, with* $0<\underline{\zeta}<\zeta<1<\underline{\zeta}(2\mathsf{k}+1)$*.*

By properly adapting the results of Wu and Zaffaroni (2018) and Zhang and Wu (2021) to the intermediate spikiness-sparsity regimes described in Section 3, we prove uniform consistency over frequencies of the smoothed periodogram pre-estimator (3).

Lemma 5.1. *Under Assumptions 3.1, 3.2, and 5.1, there exists* $G,G'>0$ *and* $d>4$*, independent of* p *and* T*, such that, as* $T\to\infty$ *and for all* $p\in\mathbb{N}$*, for* $\theta_h=\frac{h\pi}{M_T}$*,* $|h|\le M_T$*:*

1. $\mathcal{P}\left(\max_{|h|\le M_T}\frac{1}{p^\alpha}\|\widetilde{\Sigma}(\theta_h)-\Sigma(\theta_h)\|_2\le G\sqrt{\frac{M_T\log M_T}{T}}\right)\to 1$*;*
2. $\mathcal{P}\left(\max_{|h|\le M_T}\|\widetilde{\Sigma}(\theta_h)-\Sigma(\theta_h)\|_\infty\le G'\max\left(\frac{M_T p^{2/d}\log^{7/2}p}{T^{1-2/d}},\sqrt{\frac{M_T\log(M_T p)}{T}}\right)\right)\to 1.$

The proof of part 1 is new to this paper and it is a non-trivial generalization to the case of weakly pervasive factors of the result proved in Forni et al. (2017, Proposition 6) derived for the GDFM when $\alpha=1$. Part 2 is derived by adapting the results of Zhang and Wu (2021, Proposition 4.3) to the present context.

Remark 5.1. Notice that in part 1 of Lemma 5.1 the bias term, which is of order $\frac{1}{M_T^{\mathsf{k}}}$, is not included, since, for all M_T satisfying Assumption 5.1(iv), this term is always dominated by the variance term. Indeed, while the optimal choice balancing variance and squared bias is $M_T=O(T^{1/(2\mathsf{k}+1)})$, we are instead assuming $M_T=O(T^{\zeta})$ with $\zeta>\underline{\zeta}>\frac{1}{2\mathsf{k}+1}$, and with this choice of M_T the mean squared error of the smoothed periodogram is dominated by the variance, while the squared bias becomes negligible, as $T\to\infty$. Typical values of k are 1 if we choose the Bartlett kernel, or 2 if we choose the Parzen kernel.

The following theorem, proving consistency of ALSE, is the main result of the paper.

Theorem 5.1. *Define* $\varphi_{p,T,d}=\max\left(\frac{M_T p^{2/d}\log^{7/2}p}{T^{1-2/d}},\sqrt{\frac{M_T\log(M_T p)}{T}}\right)$*, with* $d=\max(d_u,d_e)>4$ *and* $\psi_0=\frac{\psi}{p^\alpha}$*, and define* $m_p=\sup_{\theta\in[-\pi,\pi]}\|S(\theta)\|_{0,v}$*. Set* $\psi_0=\frac{\varphi_{p,T,d}}{\xi(\mathcal{T})}$ *and* $\gamma\in\left[9\xi(\mathcal{T}),\frac{1}{6\mu(\Omega)}\right]$*. Suppose that Assumptions 3.1 through 3.6 and 5.1 hold and suppose also that for all* $\theta\in[-\pi,\pi]$ *there exist* $G_2,G_3>0$ *such that: (I)* $\lambda_r(L(\theta))>G_2\frac{\psi_0}{\xi^2(\mathcal{T})}$*; (II)* $\|S(\theta)\|_{min,off}>G_3\frac{\psi_0}{\mu(\Omega)}$*. Then, there exist positive reals* κ_1 *and* κ_2 *independent of* p *and* T *such that, as* $T\to\infty$ *and for all* $p\in\mathbb{N}$*, the following hold:*

1. $\mathcal{P}\left(\max_{|h|\le M_T} \frac{1}{p^{\alpha+\delta}}\|\widehat{L}(\theta_h) - L(\theta_h)\|_2 \le \kappa_1 \varphi_{p,T,d}\right) \to 1$;
2. $\mathcal{P}\left(\max_{|h|\le M_T} \|\widehat{S}(\theta_h) - S(\theta_h)\|_\infty \le \kappa_2 \varphi_{p,T,d}\right) \to 1$;
3. $\mathcal{P}\left(\mathrm{rk}(\widehat{L}(\theta_h)) = r\right) \to 1$, *for all* $\theta_h = \frac{h\pi}{M_T}$, $|h| \le M_T$;
4. $\mathcal{P}\left(\mathrm{sgn}(\mathrm{Re}(\widehat{S}(\theta_h)_{ij}) = \mathrm{sgn}(\mathrm{Re}(S(\theta_h)_{ij}))\right) \to 1$, *for all* $\theta_h = \frac{h\pi}{M_T}$, $|h| \le M_T$ *and all* $i, j = 1, \ldots, p$;
5. $\mathcal{P}\left(\max_{|h|\le M_T} \|\widehat{S}(\theta_h) - S(\theta_h)\|_2 \le \kappa_2 m_p \varphi_{p,T,d}\right) \to 1$;
6. $\mathcal{P}\left(\max_{|h|\le M_T} \frac{1}{p^{\alpha+\delta}}\|\widehat{\Sigma}(\theta_h) - \Sigma(\theta_h)\|_2 \le \kappa_1 \varphi_{p,T,d}\right) \to 1$.

This theorem is a generalization of the results by Chandrasekaran et al. (2012) for the covariance matrix of independent data having a low-rank plus sparse structure. In particular, here we go one step further by considering the case of spectral densities with r ultra-spiking eigenvalues in the sense the they increase with p^α, $\alpha \in (\frac{1}{2}, 1]$, as prescribed by Assumption 3.4, and where the residual component has a number of non-zeros also growing with p^δ, $\delta \in [0, \frac{1}{2}]$, as prescribed by Assumption 3.5.

The following remarks provide more intuitions about our results and a comparison with the existing literature.

Remark 5.2. While under the assumptions of Chandrasekaran et al. (2012) the eigen-gap does not depend on p, ours is widening as p increases. This latter assumption, which is standard in GDFM literature (Forni et al., 2017) and other factor model works (Fan et al., 2013), makes identification and thus disentangling of the low rank and sparse component easier, but on the other hand it implies convergence rates that depend polynomially on p, which appear in the rescaling terms in parts 1 and 6, and are discussed in the next remarks.

Remark 5.3. Because of the spiking behavior of the r eigenvalues of $L(\theta)$, it is natural to work with the minimization (11) rescaled by p^α. This implies that we obtain convergence rates for $\widehat{L}(\theta)$ rescaled by p^α. This is standard in this kind of literature, see e.g. Fan et al. (2013, Theorem 3) for the case of covariance matrices. This explains also the reason why we give conditions for $\psi_0 = \frac{\psi}{p^\alpha}$ rather than for ψ.

Remark 5.4. The choice of ψ_0 to be increasing with p^δ and of γ to be decreasing with p^δ, because of the definitions of $\xi(\mathcal{T})$ and $\mu(\Omega)$ in Assumption 3.6, implies that $\rho_0 = \gamma\psi_0$ does not depend on δ. This choice makes sense, indeed the sparsity of $S(\theta)$ is at most $O(p^\delta)$ and, therefore, the less sparse is $S(\theta)$

(higher δ) the more this component will become important and the more weight we need to attach to the penalty $\|L(\theta)\|_2$ in (11). Such penalization scheme gives rise to the term p^δ in parts 1 and 6. The rates of convergence are slower with respect to those in Lemma 5.1 obtained for the pre-estimator $\widetilde{\Sigma}(\theta)$. This is the price to pay for considering a penalized estimator.

Related to this, it is also useful to notice that parts 1 and 2 could be restated as follows: $\mathcal{P}\left(\max_{|h|\le M_T}\frac{1}{p^\alpha}\|\widehat{L}(\theta_h)-L(\theta_h)\|_2\le\kappa\psi_0\right)\to 1$ and $\mathcal{P}\left(\max_{|h|\le M_T}\|\widehat{S}(\theta_h)-S(\theta_h)\|_\infty\le\kappa\rho_0\right)\to 1$, respectively. In this way the role of the penalty constants in determining the convergence becomes clear. The specific definitions of ψ_0 and γ, and thus of ρ_0 are not ad hoc, but they are derived by extending the results of Chandrasekaran et al. (2012) to our setting.

Remark 5.5. Part 5 shows that the estimation error of $\widehat{S}(\theta)$ depends on the sparsity of $S(\theta)$ as measured by m_p. The smaller m_p is (more sparse), the smaller the estimation error is. By noticing that, because of Assumptions 3.3 and 3.5, it holds that $m_p\le\delta_2 p^\delta$, this result could also be restated as $\mathcal{P}\left(\max_{|h|\le M_T}\frac{1}{p^\delta}\|\widehat{S}(\theta_h)-S(\theta_h)\|_2\le\kappa_2\delta_2\varphi_{p,T,d}\right)\to 1$, which represents the worst case scenario in which $S(\theta)$ is the least sparse compatibly with Assumption 3.5. Notice also that, since $\delta<\alpha$, the effect of sparsity in part 6 is always dominated by the estimation error of the low rank component $\widehat{L}(\theta)$.

Remark 5.6. For the estimation algorithm to work properly it is crucial that $\mathrm{rk}(L(\theta))$ is constant across frequencies as required by Assumption 3.3(i). As already mentioned this is standard in GDFM literature. Moreover, we notice that, although in principle we allow for the number of nonzero elements of $S(\theta)$ to be frequency dependent, ALSE is in fact able to control only for the maximum number of nonzeros across all frequencies. This is because, as shown in Appendix A, ALSE is equivalent to looking for a solution of the minimization (11) restricted to Ω, which depends only on S_∞ and in turn it holds that $\sup_{\theta\in[-\pi,\pi]}\|S(\theta)\|_{0,v}=\|S_\infty\|_{0,v}$.

By looking at the rate $\varphi_{p,T,d}$, it is clear that in general p can grow at most polynomially in T, so let us assume $p\asymp T^\eta$ for some $\eta>0$. Let us also consider the standard case in which we use the Bartlett kernel with bandwidth to $M_T=\lfloor\sqrt{T}\rfloor$ to compute the pre-estimator, i.e., we set $\zeta=\frac{1}{2}$ and $\mathsf{k}=1$ in Assumption 5.1. These are the same choices used in Section 8 and in the simulation study. Furthermore, recall that d denotes the minimum number of moments of $\{u_t\}$ and $\{e_t\}$ we require to exist, which by

Assumption 3.1 must be such that $d > 4$.

Clearly, under these conditions, the second term in $\varphi_{p,T,d}$ is always decreasing as $T \to \infty$. A necessary condition for the first term to converge is $d > 2(1+\eta)$, which is non binding as long as $\eta \le 1$ since, by assumption, $d > 4$. However, if p increases faster than T, i.e., $\eta > 1$, we must guarantee the existence of more moments to still have consistency. If, for example, $\eta = 2$, then we must have $d > 6$. Summing up, in general, the larger is p the lighter the tails have to be in order to guarantee consistency.

As far as the rate of consistency is concerned, we notice that if $d > 8(1+\eta)$ we get the *classical* consistency rate $\varphi_{p,T,d} = \sqrt{\frac{M_T \log(M_T p)}{T}}$, which is similar to results for the estimation of large covariance matrices, see also Remark 5.7 below. In particular, this rate is achieved, regardless of η, in the sub-Gaussian case, i.e., when we can set $d = \infty$, in which case we can also allow for larger values of p as long as $\log p = o(T)$. However, in general, when we are in presence of heavy tails, we would typically have the slower consistency rate $\varphi_{p,T,d} = \frac{M_T p^{2/d} \log^{7/2} p}{T^{1-2/d}}$, unless p is very small.

To have consistency we also need conditions (I) and (II) to hold. Condition (I) must be compatible with Assumption 3.4, which requires $\lambda_r(L(\theta)) \asymp p^\alpha$ for all $\theta \in [-\pi, \pi]$. Thus, we must require that $\frac{\psi_0}{\xi^2(T)p^\alpha} \to 0$, as $p, T \to \infty$. In light of the previous comments, setting again $M_T = \lfloor \sqrt{T} \rfloor$, we must have either $\frac{p^{6\delta - 2\alpha + 4/d}}{T} \to 0$ or $\frac{p^{12\delta - 4\alpha}}{T} \to 0$, in the heavy tail or in the sub-Gaussian case, respectively. Both conditions are always true when $\delta = 0$, while they require larger T the smaller the eigengap gets (decreasing $\alpha - \delta$). A sufficient condition for both to hold is $\delta \le \frac{\alpha}{3} - \frac{1}{6} + \epsilon$, for some $\epsilon > 0$.

Condition (II) must be compatible with the obvious requirement $\|S(\theta)\|_{min,off} \le G_3'$ for some $G_3' > 0$. This is always true, indeed, by Assumption 3.6, because condition (II) can be written as $\|S(\theta)\|_{min,off} > G_3 \frac{\kappa_s}{\sqrt{r \kappa_L}} \varphi_{p,T,d}$, which under the conditions on p, T, and d, stated above, is always decreasing to zero as $p, T \to \infty$.

The following Corollary characterizes the inverse of the estimated spectral density matrix.

Corollary 5.1. *Suppose that all assumptions of Theorem 5.1 hold, and suppose also that* $m_p \varphi_{p,T,d} \to 0$, *as* $T \to \infty$ *and for all* $p \in \mathbb{N}$. *Then, there exist positive reals* κ_3 *and* κ_4 *independent of* p *and* T *such that, as* $T \to \infty$ *and for all* $p \in \mathbb{N}$, *the following hold:*

1. $\mathcal{P}\left(\min_{|h| \le M_T} \lambda_p(\widehat{S}(\theta_h)) \ge \kappa_3\right) \to 1$;
2. $\mathcal{P}\left(\max_{|h| \le M_T} \|\widehat{S}^{-1}(\theta_h) - S^{-1}(\theta_h)\|_2 \le \kappa_4 m_p \varphi_{p,T,d}\right) \to 1$;

3. $\mathcal{P}\left(\min_{|h|\le M_T} \lambda_p(\widehat{\Sigma}(\theta_h)) \ge \kappa_3\right) \to 1.$

If we further suppose that $p^{\alpha+\delta} m_p \varphi_{p,T,d} \to 0$*, then, there exists a positive real* κ_5 *independent of* p *and* T *such that, as* $T \to \infty$ *and for all* $p \in \mathbb{N}$*, the following hold:*

4. $\mathcal{P}\left(\max_{|h|\le M_T} \frac{1}{p^{\alpha+\delta}} \|\widehat{\Sigma}^{-1}(\theta_h) - \Sigma^{-1}(\theta_h)\|_2 \le \kappa_5 \varphi_{p,T,d}\right) \to 1.$

Remark 5.7. The results in Theorem 5.1 and Corollary 5.1 can be compared to those obtained by Fan et al. (2013, Remark 2) for the estimation of a large covariance matrix with a low rank component generated by r strong factors, plus a sparse component having bounded eigenvalues and $\ell_{0,v}$ norm equal to m_p (by considering the exact sparse case in their work, i.e., when setting $q = 0$ therein). They also assume to observe time series drawn from a distribution with sub-exponential tails, thus assuming all moments to exist, and obtain for their estimator of the sparse component and its inverse a consistency rate $O_p\left(m_p\sqrt{\frac{\log p}{T}}\right)$. This is also the same rate obtained by Bickel and Levina (2008, Theorem 2) when considering the case of a large purely sparse covariance matrix with Gaussian entries. Using our notation the assumptions of Fan et al. (2013) correspond to setting $\alpha = 1$, $\delta = 0$, and $d = \infty$, in which case from part 5 of Theorem 5.1 and from part 2 of Corollary 5.1, we would get a similar rate, which, with our notation, is $O_p\left(m_p\sqrt{\frac{M_T \log(M_T p)}{T}}\right)$. Concerning the estimator of the entire covariance matrix, Fan et al. (2013, Remark 3) obtain a consistency rate $O_p\left(p\, m_p\sqrt{\frac{\log p}{T}}\right)$, which is similar to the rate $O_p\left(p\, m_p\sqrt{\frac{M_T \log(M_T p)}{T}}\right)$ that what we would get from part 6 of Theorem 5.1 when $\alpha = 1$, $\delta = 0$, and $d = \infty$.

Remark 5.8. Concerning the estimator of the inverse spectral density in part 4 of Corollary 5.1, we obtain the same rate of the estimator of $\Sigma(\theta)$ given in part 6 of Theorem 5.1, which is worse than what obtained for the inverse covariance matrix by Fan et al. (2013). To obtain a comparable rate we should follow the approach of Fan et al. (2013, Equation (2.13) and Appendix C.4.2) and define the inverse estimator by explicitly exploiting the existence of a low-rank component and then applying the Sherman-Morrison-Woodbury formula. This, however, would require to explicitly estimate also the coefficients B_s of the common filters (see (5)), which is beyond the scope of this paper, and, therefore, it is left for further research.

6 Unshrinking

ALSE may suffer from systematic sub-optimality for what concerns the estimated eigenvalues. Indeed, if p is large and the latent eigenvalues are spiked ($\alpha \simeq 1$), the singular value thresholding procedure by Cai et al. (2010) may lead to the over-shrinkage of the eigenvalues of $\widehat{L}(\theta_h)$. For this reason, following Farnè and Montanari (2020), we propose to un-shrink those eigenvalues. The new estimate of $S(\theta)$ is then obtained by keeping fixed the off-diagonal sparsity pattern recovered, and deriving its diagonal by difference from the diagonal of $\widehat{\Sigma}(\theta_h)$. More specifically, let $\widehat{r} = \mathrm{rk}(\widehat{L}(\theta_h))$ and consider the spectral decomposition $\widehat{L}(\theta_h) = \widehat{W}(\theta_h)\widehat{D}(\theta_h)\widehat{W}^{\dagger}(\theta_h)$, where $\widehat{D}(\theta_h)$ is $\widehat{r} \times \widehat{r}$ diagonal matrix of the nonzero eigenvalues of $\widehat{L}(\theta_h)$, and $\widehat{W}(\theta_h)$ is the $p \times \widehat{r}$ matrix of corresponding normalized eigenvectors. Then, for any $\theta_h = \frac{h\pi}{M_T}$, $|h| \leq M_T$, the resulting UNshrunk ALgebraic Spectral Estimator (UNALSE) is defined as:

$$
\begin{aligned}
&\widehat{L}_{\text{UNALSE}}(\theta_h) = \widehat{W}(\theta_h)\left(\widehat{D}(\theta_h) + \psi I_r\right)\widehat{W}^{\dagger}(\theta_h),\\
&\mathrm{diag}(\widehat{S}_{\text{UNALSE}}(\theta_h)) = \mathrm{diag}(\widehat{\Sigma}(\theta_h)) - \mathrm{diag}(\widehat{L}_{\text{UNALSE}}(\theta_h)), \quad \text{off-diag}(\widehat{S}_{\text{UNALSE}}(\theta_h)) = \text{off-diag}(\widehat{S}(\theta_h)),\\
&\widehat{\Sigma}_{\text{UNALSE}}(\theta_h) = \widehat{L}_{\text{UNALSE}}(\theta_h) + \widehat{S}_{\text{UNALSE}}(\theta_h).
\end{aligned}
$$

By construction $\mathrm{rk}(\widehat{L}_{\text{UNALSE}}(\theta_h)) = \mathrm{rk}(\widehat{L}(\theta_h))$ and $\mathrm{supp}(\widehat{S}_{\text{UNALSE}}(\theta_h)) = \mathrm{supp}(\widehat{S}(\theta_h))$, thus the algebraic consistency of ALSE, proved in parts 3 and 4 of Theorem 5.1, is preserved by UNALSE. Theorem A.1 in Appendix A also shows that, under the same assumptions and conditions of Theorem 5.1, UNALSE is the closest (according to the Frobenius norm) estimator to the smoothed periodogram pre-estimator.

7 Threshold selection

In solving problem (11), the choice of the thresholds ψ and ρ is a nontrivial issue. Let us denote the solutions of (11) with given thresholds ψ and ρ as $\widehat{L}_{\psi,\rho}(\theta_h)$, $\widehat{S}_{\psi,\rho}(\theta_h)$, and $\widehat{\Sigma}_{\psi,\rho}(\theta_h) = \widehat{L}_{\psi,\rho}(\theta_h) + \widehat{S}_{\psi,\rho}(\theta_h)$. For any $\theta_h = \frac{h\pi}{M_T}$, $|h| \leq M_T$, we consider the following criterion:

$$
MC_h(\psi, \rho) = \max\left\{\frac{\widehat{r}_{\psi,\rho}\|\widehat{L}_{\psi,\rho}(\theta_h)\|_2}{\psi\widehat{\beta}_{\psi,\rho}(\theta_h)}, \frac{\|\widehat{S}_{\psi,\rho}(\theta_h)\|_{1,v}}{\rho(1-\widehat{\beta}_{\psi,\rho}(\theta_h))}\right\}, \tag{13}
$$

where $\widehat{r}_{\psi,\rho} = \mathrm{rk}(\widehat{L}_{\psi,\rho}(\theta_h))$ and $\widehat{\beta}_{\psi,\rho}(\theta_h) = \frac{\mathrm{tr}(\widehat{L}_{\psi,\rho}(\theta_h))}{\mathrm{tr}(\widehat{\Sigma}_{\psi,\rho}(\theta_h))}$ is the estimated proportion of latent variance. The optimal threshold pair $(\breve{\psi}_h, \breve{\rho}_h)$ is thus selected as the mini-max $(\breve{\psi}_h, \breve{\rho}_h) = \arg\min_{\psi,\rho,\widehat{S}_{\psi,\rho}(\theta_h)\succ 0} MC_h(\psi,\rho)$, where ψ and ρ vary across pre-specified grids and the minimum is taken over the threshold pairs that return a positive definite $\widehat{S}_{\psi,\rho}(\theta_h)$. This threshold selection method penalizes solution pairs with too dispersed latent eigenvalues and too many residual nonzeros in single rows, by comparing two appropriately re-scaled versions of the spectral norm of the low rank solution and the row-wise maximum norm of the residual solution.

In order to ensure the effectiveness of the above approach, the threshold grids need to be properly chosen at each frequency θ_h. From the conditions in Theorem 5.1, we must always have $\psi > \frac{1}{\xi(\mathcal{T})}\sqrt{\frac{1}{T}}$. Then, we recall from (Chandrasekaran et al., 2011, Proposition 4) that we must have (see also (9)) $\mathrm{inc}(L(\theta_h)) \leq \xi(\mathcal{T}) \leq 2\,\mathrm{inc}(L(\theta_h))$ with $\sqrt{\frac{r}{p}} \leq \mathrm{inc}(L(\theta_h)) \leq 1$, where $\mathrm{inc}(L(\theta_h))$ is the incoherence of $L(\theta_h)$, defined as $\mathrm{inc}(L(\theta_h)) = \max_{i=1,\dots,p} \|\mathcal{P}e_i\|_2$, with e_i the canonical basis vector (ith column of the p-dimensional identity matrix), and the operator $\mathcal{P}$ projecting each e_i onto the row/column space of $L(\theta_h)$. In light of this, we initialize the grid for the threshold ψ as the sequence of n_{thr} equi-spaced real numbers from $\sqrt{\frac{p}{T}}\frac{1}{2\widetilde{\mathrm{inc}}}$ to $\sqrt{\frac{p}{T}}\frac{1}{\widetilde{\mathrm{inc}}}$, where, for any given positive real r_{thr}, we set $\widetilde{\mathrm{inc}} = \sqrt[4]{\frac{r_{thr}}{p}}$, which is the geometric mean of the minimum and maximum incoherence values.

Concerning the sparsity threshold, we let $\rho = \rho_1\sqrt{\frac{p}{T}}\frac{1}{\widetilde{\mathrm{inc}}}$ and we initialize the grid for the threshold ρ_1 as the sequence of n_{thr} equi-spaced real numbers from $\frac{s_{thr}}{\sqrt{p}}$ to $\frac{s_{thr}}{\sqrt[4]{p}}$, where s_{thr} is a given positive real. The two extremes represent two plausible extreme values for the proportion of nonzero entries satisfying the assumptions of Theorem 5.1. Once we have a grid for ρ_1 we immediately have a grid for ρ.

In practice, we set $n_{thr} = 10$ and we adopt the following recursive approach.

1. Set $s_{thr} = 1$ and increase it until we get thresholds $(\breve{\psi}_{h,1}, \breve{\rho}_{h,1})$ such that the ALSE $\widehat{S}_{\breve{\psi}_{h,1},\breve{\rho}_{h,1}}(\theta_h)$ is diagonal for all frequencies; denote as $s_{thr,1}$ the chosen value of $s_{thr,1}$.
2. Set $r_{thr} = 1$ and increase it until we get thresholds $(\breve{\psi}_{h,2}, \breve{\rho}_{h,2})$ such that $\breve{\psi}_{h,2}$ is close but not equal to the left-extreme of the grid and the ALSE $\widehat{L}_{\breve{\psi}_{h,2},\breve{\rho}_{h,2}}(\theta_h)$ has rank constant across frequencies.
3. Starting from $s_{thr} = s_{thr,1}$ decrease s_{thr} until we get thresholds $(\breve{\psi}_{h,3}, \breve{\rho}_{h,3})$ such that the ALSE $\widehat{S}_{\breve{\psi}_{h,3},\breve{\rho}_{h,3}}(\theta_h)$ is non-diagonal, but positive definite for all frequencies, and the selected $\breve{\rho}_{h,3}$ not equal to the left-extreme of the grid.

4. The optimal thresholds are given by $\breve{\psi}_h = \breve{\psi}_{h,2}$ and $\breve{\rho}_h = \breve{\rho}_{h,3}$.

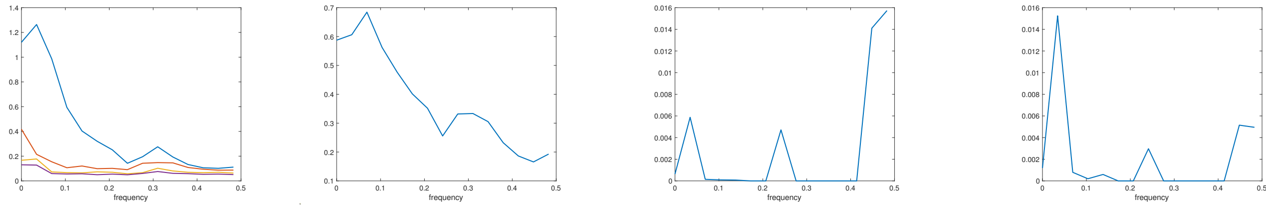
8 Real data analysis

We consider a dataset of $p = 101$ quarterly macroeconomic indicators regarding the US economy observed over $T = 210$ time points spanning the period 1960:Q2-2012:Q3 (see, e.g., McCracken and Ng, 2020). Throughout, we compute the smoothed periodogram by setting $M_T = \lfloor \sqrt{T} \rfloor = 14$ and using the Bartlett kernel. UNALSE is computed as described in Sections 4 and 6, while the thresholds are chosen as explained in Section 7. Results are reported using frequencies $f_h = \frac{\theta_h}{2\pi} = \frac{h}{2M_T}$, $|h| \le M_T$, measured in cycles per unit of time. Thus, for a given frequency f_h the corresponding period is given by f_h^{-1} and, since our data is quarterly, its unit of measure are quarters, with one quarter being equal to three months.

In the first panel of Figure 1 we show the four largest eigenvalues, rescaled by p, of the smoothed periodogram estimator. The top eigenvalue shows a maximum at $f_h = 0.07$ and a second peak at $f_h = 0.31$, corresponding to periods of about 3.5 years and 9 months, respectively. Note that 3.5 years is around the typical period of a business cycle, while the higher frequency peak is typically more related to nominal variables such as inflation (see also the results in Barigozzi and Luciani, 2021, and below). The estimated rank by UNALSE is $\widehat{r} = 2$ at all frequencies, which is an agreement with the recent findings by Avarucci et al. (2022), who interpret the two common factors as a demand (high-frequency) and a supply (low-frequency) factors (see also Angeletos et al., 2020). The second panel of Figure 1 shows the proportion of latent variance $\widehat{\beta}(f_h) = \frac{\mathrm{tr}(\widehat{L}(f_h))}{\mathrm{tr}(\widehat{\Sigma}(f_h))}$ so the contribution of $\widehat{L}(f_h)$, which follows the pattern of the leading eigenvalues of $\widetilde{\Sigma}(f_h)$, hence it captures the business cycle frequency. The third panel of Figure 1 reports the proportion of residual covariance $\widehat{\varepsilon}(f_h) = \frac{\sum_{i=1}^{p}\sum_{j=i+1}^{p} |\widehat{S}_{ij}(f_h)|}{\sum_{i=1}^{p}\sum_{j=i+1}^{p} |\widehat{\Sigma}_{ij}(f_h)|}$, summarizing the contribution of $\widehat{S}(f_h)$, which is maximum at frequency $f_h = 0.48$, corresponding to a period of about 6 months. Secondary maxima are at $f_h = 0.03$, i.e., a period of 8 years, and $f_h = 0.24$, corresponding to a period of 1 year. Finally, the fraction of nonzeros in the fourth panel of Figure 1 has a similar pattern.

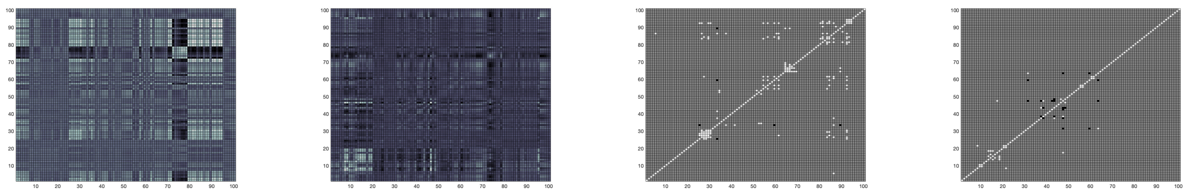
In Figure 2 (first two panels) we show heat-maps of $\widehat{L}(f_h)$ at frequencies 0.07 and 0.31. The variables contributing more to the comovements at the low frequency ($f_h = 0.07$) are either related to the labor market (variables from $i = 70, \ldots, 94$) such as (i) Unemployment rate; (ii) Average Mean Duration

Figure 1: US macroeconomic data - Co-movements and sparsity. From the left, we report over the chosen frequencies $f_h = \frac{h}{29}$, $h = 0, \ldots, 14$, the four estimated eigenvalues, the estimated latent variance proportion, the estimated residual covariance proportion, and the fraction of residual nonzeros.



$\frac{\widehat{\lambda}_j(\widetilde{\Sigma}(f_h))}{p}$, $j = 1, 2, 3, 4.$ $\quad$ $\widehat{\beta}(f_h) = \frac{\mathrm{tr}(\widehat{L}(f_h))}{\mathrm{tr}(\widehat{\Sigma}(f_h))}$. $\quad$ $\widehat{\varepsilon}(f_h) = \frac{\sum_{i=1}^{p}\sum_{j=i+1}^{p}|\widehat{S}_{ij}(f_h)|}{\sum_{i=1}^{p}\sum_{j=i+1}^{p}|\widehat{\Sigma}_{ij}(f_h)|}$. $\quad$ fraction of nonzeros in $\widehat{S}(f_h)$.

Figure 2: US macroeconomic data - Heat maps of $\mathrm{Re}(\widehat{L}(f_h))$ and $\mathrm{Re}(\widehat{S}(f_h))$ at selected frequencies. The X-axis and the Y-axis contain the 101 variables of the dataset. Positive values are in white, negative ones in black, zeros in gray.



$\mathrm{Re}(\widehat{L}(f_h = 0.07))$. $\quad$ $\mathrm{Re}(\widehat{L}(f_h = 0.31))$. $\quad$ $\mathrm{Re}(\widehat{S}(f_h = 0.03))$. $\quad$ $\mathrm{Re}(\widehat{S}(f_h = 0.48))$.

of Unemployment; (iii) All Employees in various sectors, or are related to credit market and the real economy (variables from $i = 25, \ldots, 37$) such as GDP, Investment, and Consumption growth rates. All these are the variables typically driving the business cycle. Note that these variables drive also the comovements at $f_h = 0$. At higher frequency ($f_h = 0.31$) the most relevant variables are the growth rates, i.e., inflation, of Consumer and Production Price Indexes (variables from $i = 7, \ldots, 19$). Thus the two main sources of common variation are related to: (i) the real economy, and (ii) to the nominal economy.

In Figure 2 (last two panels) we show heat-maps of $\widehat{S}(f_h)$ at frequencies 0.03 and 0.48. It is worth mentioning some prominent co-spectral relationship at selected frequencies. At frequency $f_h = 0.03$, the following pairs display strong co-dependence: (i) Consumer Loans at All Commercial Banks and Total Consumer Credit (Owned and Securitized); (ii) 3-Year Treasury Constant Maturity Rate and 10-Year Treasury Constant Maturity Rate; (iii) Compensation Per Hour in the Manufacturing Sector and in the Business Sector. At frequency $f_h = 0.48$, we observe a strong relationship for the following variable pairs:

(i) Consumer Price Index for All Urban Consumers of All Items Less Energy and of All Items Less Food & Energy; (ii) Real Imports and Exports of Goods & Services Per-Capita; (iii) Real Exports of Goods & Services and Gross Private Domestic Investment Price Index; (iv) Unit Labor Cost in the Business Sector and Output Per Hour of All Persons in the Business Sector. At frequency $f_h = 0.24$, the strongest co-spectral relationships are instead (not shown): (i) Producer Price Index of Industrial Commodities and of All Commodities; (ii) Producer Price Index of Industrial Commodities and Treasury Constant Maturity rate. These local, co-movements can be interpreted as due to weaker, idiosyncratic, factors related to the banking system represented by credit variables (low-frequency), to prices and interest rates, i.e., related to monetary policy (mid-frequency), and to the trade/consumption dimension of the US economy (high-frequency). No relevant idiosyncratic comovements are found in the long run, i.e., at $f_h = 0$.

9 Conclusions

In this paper, we consistently estimate the spectral density matrix under the assumption of a low rank plus sparse multivariate spectrum for the data. We prove that the nuclear norm plus l_1 norm heuristics consistently recovers across frequencies the spectral components and their sum, as well as the dynamic rank and the residual sparsity pattern. We call the resulting estimators UNALSE (UNshrunk ALgebraic Spectral Estimator). The empirical implications of our approach are discussed on a standard US macroeconomic dataset, showing that we are able to catch the driving variables of the latent dynamics, as well as the particular strength of specific local relationships which might be heterogeneous across frequencies. This opens up the way to enhanced dynamic factor scores estimation and temporal network analysis.

Acknowledgements and Declarations

Matteo Barigozzi gratefully acknowledges the support by MIUR (PRIN 2020) under Grant 2020N9YFFE. Matteo Farnè gratefully acknowledges the support by the *Accademia Nazionale dei Lincei* under Grant *British Academy 2018*.

The authors report there are no competing interests to declare.

References

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., and Veronese, G. (2010). New Eurocoin: Tracking economic growth in real time. *The Review of Economics and Statistics*, 92(4):1024–1034.

Angeletos, G.-M., Collard, F., and Dellas, H. (2020). Business-cycle anatomy. *American Economic Review*, 110(10):3030–70.

Avarucci, M., Cavicchioli, M., Forni, M., and Zaffaroni, P. (2022). The main business cycle shock (s): Frequency-band estimation of the number of dynamic factors. Technical Report No. DP17281, CEPR.

Barigozzi, M. and Hallin, M. (2017). A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):581–605.

Barigozzi, M. and Luciani, M. (2021). Measuring the output gap using large datasets. *The Review of Economics and Statistics*. available online.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Böhm, H. and von Sachs, R. (2008). Structural shrinkage of nonparametric spectral estimators for multivariate time series. *Electronic Journal of Statistics*, 2:696–721.

Böhm, H. and von Sachs, R. (2009). Shrinkage estimation in the frequency domain of multivariate time series. *Journal of Multivariate Analysis*, 100(5):913–935.

Breitung, J. and Candelon, B. (2006). Testing for short-and long-run causality: A frequency-domain approach. *Journal of Econometrics*, 132(2):363–378.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory*. SIAM.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982.

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

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596.

Chaudhuri, S. E. and Lo, A. W. (2015). Spectral analysis of stock-return volatility, correlation, and beta. In *2015 IEEE Signal Processing and Signal Processing Education Workshop*, pages 232–236.

Corbae, D., Ouliaris, S., and Phillips, P. C. (2002). Band spectral regression with trending data. *Econometrica*, 70(3):1067–1109.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172.

Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457.

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal l_1 norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59:797–829.

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Farnè, M. and Montanari, A. (2020). A large covariance matrix estimator under intermediate spikiness regimes. *Journal of Multivariate Analysis*, 176:104577.

Farnè, M. and Montanari, A. (2021). A bootstrap method to test Granger-causality in the frequency domain. *Computational Economics*.

Fazel, M., Hindi, H., and Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE.

Fiecas, M., Leng, C., Liu, W., and Yu, Y. (2019). Spectral analysis of high-dimensional time series. *Electronic Journal of Statistics*, 13(2):4079–4101.

Fiecas, M. and Ombao, H. (2011). The generalized shrinkage estimator for the analysis of functional connectivity of brain signals. *The Annals of Applied Statistics*, 5(2A):1102–1125.

Fiecas, M. and Ombao, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association*, 111(516):1440–1453.

Fiecas, M. and von Sachs, R. (2014). Data-driven shrinkage of the spectral density matrix of a high-dimensional time series. *Electronic Journal of Statistics*, 8(2):2975–3003.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82(4):540–554.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471):830–840.

Forni, M., Hallin, M., Lippi, M., and Zaffaroni, P. (2017). Dynamic factor models with infinite-dimensional factor space: asymptotic analysis. *Journal of Econometrics*, 199(1):74–92.

Forni, M. and Lippi, M. (2001). The generalized dynamic factor model: representation theory. *Econometric theory*, 17(6):1113–1141.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89:2409–2437.

Hallin, M. and Lippi, M. (2013). Factor models in high-dimensional time series. A time-domain approach. *Stochastic Processes and their Applications*, 123(7):2678–2695.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617.

Harvey, A. C. (1978). Linear regression in the frequency domain. *International Economic Review*, 19(2):507–512.

Joyeux, R. (1992). Tests for seasonal cointegration using principal components. *Journal of Time Series Analysis*, 13(2):109–118.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254.

Luo, X. (2011). Recovering model structures from large low rank and sparse covariance matrix estimation. *arXiv:1111.1133*.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457.

McCracken, M. and Ng, S. (2020). FRED-QD: A quarterly database for macroeconomic research. Technical Report 26872, National Bureau of Economic Research.

Müller, U. K. and Watson, M. W. (2018). Long-run covariability. *Econometrica*, 86(3):775–804.

Ombao, H., Von Sachs, R., and Guo, W. (2005). SLEX analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100(470):519–531.

Ombao, H. C., Raz, J. A., von Sachs, R., and Malow, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, 96(454):543–560.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479.

Sargent, T. and Sims, C. (1977). Business cycle modeling without pretending to have too much a priori economic theory. Technical report, Federal Reserve Bank of Minneapolis.

Stock, J. H. and Watson, M. W. (1988). Testing for common trends. *Journal of the American Statistical Association*, 83(404):1097–1107.

Velasco, C. and Robinson, P. M. (2000). Whittle pseudo-maximum likelihood estimation for nonstationary time series. *Journal of the American Statistical Association*, 95(452):1229–1243.

Wu, W. B. and Zaffaroni, P. (2018). Asymptotic theory for spectral density estimates of general multivariate time series. *Econometric Theory*, 34:1–22.

Zhang, D. and Wu, W. B. (2021). Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *The Annals of Statistics*, 49(1):233–254.