On Factor Models for High-Dimensional Time Series

Ragvir Singh Sabharwal

Department of Statistics

The London School of Economics and Political Science

A thesis submitted for the degree of

Doctor of Philosophy

London, March 2023

The magic of **Char**
The mystery of **Coal**
The mischief of **Slate**
The mayhem of **Marble**
The marvel of **Soxy**
The memory of **Bouncy**

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science (LSE) is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

Statement of conjoint work:

I acknowledge that Chapter 2, albeit fully authored by myself, benefitted significantly from guidance by Professor Matteo Barigozzi. I estimate my own contribution towards this work to be around 90%.

Ragvir Singh Sabharwal
March 2023

Acknowledgements

I first acknowledge Professor Barigozzi, without whom my thesis would have remained unwritten. He was ever the bamboo to my climbing rose! I am also deeply grateful to my examiners, Professor Trapani and Dr Chen, for engaging with my work and helping me improve it.

I thank my supervisor Professor Moustaki and other members of the Statistics faculty as well as Penny Montague and other members of the Statistics administration. I cherish the professional relationships I have been able to develop with such esteemed colleagues.

I acknowledge my PhD colleagues with an extra special nod to those with whom I previously shared office spaces. Their names, too many to mention here, are etched forever in fond memory. I wish each one of them the highest success for the future.

I acknowledge the Economic and Social Research Council for providing funding for the first few years of my research. I am also grateful to the Departments of Statistics and Economics and to Dr Mavrakakis for providing so many wonderful opportunities for me to teach at LSE.

I acknowledge Mum, Dad, Fries, Aveer, Seher and my entire loving family, for their unwavering support and their belief in me. I acknowledge Alice for being an unrivalled best friend; and I encourage Tayfun to try harder to take her spot. I thank Haziq for his brotherly companionship, and Rosemary for helping me pronounce the dreaded word *thefiff*.

Finally, I acknowledge my companion underneath the mango tree, my confidante, my soulmate, my rock, and my wife. Thank you, Neha, for being as invested in my journey as I was. I love you so very much.

Abstract

The aim of this thesis is to develop statistical methods for use with factor models for high-dimensional time series. We consider three broad areas: estimation, changepoint detection, and determination of the number of factors.

In Chapter 1, we sketch the backdrop for our thesis and review key aspects of the literature.

In Chapter 2, we develop a method to estimate the factors and parameters in an approximate dynamic factor model. Specifically, we present a spectral expectation-maximisation (or "spectral EM") algorithm, whereby we derive the E and M step equations in the frequency domain. Our E step relies on the Wiener-Kolmogorov smoother, the frequency domain counterpart of the Kalman smoother, and our M step is based on maximisation of the Whittle Likelihood with respect to the parameters of the model. We initialise our procedure using dynamic principal components analysis (or "dynamic PCA"), and by leveraging results on lag-window estimators of spectral density by Wu and Zaffaroni (2018), we establish consistency-with-rates of our spectral EM estimator of the parameters and factors as both the dimension (N) and the sample size (T) go to infinity. We find rates commensurate with the literature. Finally, we conduct a simulation study to numerically validate our theoretical results.

In Chapter 3, we develop a sequential procedure to detect changepoints in an approximate static factor model. Specifically, we define a ratio of eigenvalues of the covariance matrix of N observed variables. We compute this ratio each period using a rolling window of size m over time, and declare a changepoint when its value breaches an alarm

threshold. We investigate the asymptotic behaviour (as $N, m \to \infty$) of our ratio, and prove that, for specific eigenvalues, the ratio will spike upwards when a changepoint is encountered but not otherwise. We use a block-bootstrap to obtain alarm thresholds. We present simulation results and an empirical application based on Financial Times Stock Exchange 100 Index (or "FTSE 100") data.

In Chapter 4, we conduct an exploratory analysis which aims to extend the randomised sequential procedure of Trapani (2018) into the frequency domain. Specifically, we aim to estimate the number of dynamically loaded factors by applying the test of Trapani (2018) to eigenvalues of the estimated spectral density matrix (as opposed to the covariance matrix) of the data.

Contents

"Factor model methods [in a high-dimensional context] are the ideal tool – arguably, the only successful ones."
– Barigozzi and Hallin (2015)

Chapter 1

Introduction

This thesis is concerned with the development of statistical methods for use with factor models designed for high-dimensional time series. Accordingly, in this opening chapter, we introduce the rich field of factor models and associated methods. In Section 1.1, we explore models. In particular, we (i) provide a historical overview of research in the area, (ii) discuss examples of applications of factor models, and (iii) explain the taxonomy of models considered in the literature. In this respect, our goal is not only to set the broad scene for our own study, but also to share with the reader our sense that undertaking research in this area is a fruitful endeavour. In Section 1.2, we narrow in on three sub-fields of the literature that tackle key methods. In particular, we summarise studies in the areas of estimation, changepoint detection and determination of the number of factors. We relegate any additional details about the literature not covered herein to the main bodies of respective upcoming chapters.

1.1 Synopsis of the Literature on Models

Factor analysis is a highly effective way of modelling inter-relationships between potentially large sets of observed variables in a parsimonious manner. In other words, it is an ideal tool for ushering in a reduction in dimensionality while preserving most of the covariation information in the data. The success of factor analysis in tackling the high-dimensional time series that we typically find in contemporary macroeconomics and finance belies its pre-"big data" origins. Indeed, most monographs on factor analysis begin by noting the old chestnut that research into this area goes back more than a century, originating in the field of psychology with Spearman (1904). We too begin this thesis with the same frequently-quoted fact not because we want to maintain an unexciting conformance with academic tradition but because we truly believe that this fact is testament to both the wide-ranging applicability of factor model methods and their longevity.

Accordingly, with fresh enthusiasm, we touch upon some key historical developments and provide examples of applications of factor analysis. In tandem, we use this introductory section to map out the salient features of the landscape wherein this study resides. It is no surprise that model specifications in this area may be simple, complex or anywhere in-between, as also the statistical methods employed in conjunction with them. Therefore, in the paragraphs below, we do not attempt to cover every possible variant of models encountered in the literature nor every possible methodological development. Nevertheless, we recount a generally-accepted categorisation of factor models together with brief descriptions of what each category entails, and we signpost readers looking for a richer variety of surveys of the literature to *inter alia* the works of Breitung and Eickmeier (2006), Bai and Ng (2008), Stock and Watson (2016), Bai and Wang (2016), Doz and Fuleky (2020), Barigozzi (2020) and Hallin et al. (2020).

1.1.1 Factor Model Decomposition

We start right at the wellspring of the merits of factor models. To fix ideas, let us consider models for (observed) realisations from a zero-mean double indexed

stochastic process x_{it} for cross-sectional index $i = 1, ..., N$ and temporal index $t = 1, ..., T$. Distinctions among individual specifications notwithstanding, all factor models postulate the same (unobserved) decomposition of x_{it} into a sum of two mutually orthogonal[1] components, one that is *common* to all units in the cross-section χ_{it} (thus capturing co-movements), and one that is *idiosyncratic* ξ_{it} (thus accounting for any remaining individual features). The importance of this decomposition emanates from its two critical properties. The first is the reduced rank nature of χ_{it}, which is to say that it is driven by a small number $q \ll N$ (with q independent of N) of exogenously given primitives or shocks; and the second is that the cross-sectional correlations amongst the ξ_{it} terms are sufficiently weak (failing which, idiosyncracies would be hard to distinguish from commonalities). Therein lies our inherent reduction in dimensionality.

1.1.2 Applications

The next question is whether the factor structure postulated above is reasonable in real-world scenarios. Are there convincing examples of its use beyond just the discussion by Spearman (1904) of general intelligence as the underlying unobserved driver of various measures of cognitive ability of an individual?

Indeed, several authors as far back as Burns and Mitchell (1946) have considered an unobserved business cycle as the driver of co-movements among macroeconomic aggregates. Engle and Watson (1981) considers an unobserved local metropolitan-area wage rate as the driver of wages across sectors throughout the United States (US). Diebold et al. (2006) considers *inter alia* unobserved yield-curve factors (level, slope, curvature) as the drivers of bond yields. Mody and Taylor (2007) considers an unobserved regional vulnerability as the driver of country-specific measures of financial stress. Barigozzi and Hallin (2015) considers an unobserved market volatility as the driver of individual asset liquidity measures. The list could go on but we stop here. Our point is simply, as phrased by Doz and Fuleky (2020, p.418), that "[...] empirical evidence supports their main premise: [factor models] fit the data". Nevertheless, even if one were per-

[1] That is, uncorrelated at all leads and lags

suaded that applications abound, one could still ask what one can do with the results of factor analysis. Once again, the literature is replete with examples and we mention a handful below.

A first arena is that of forecasting and factor methods are routinely deployed by central banks and research institutes for this purpose. References include Stock and Watson (2002a,b) which predict real and nominal macroeconomic variables in the US; and Forni et al. (2003) and Banerjee et al. (2005) in the euro area. Relatedly, factor methods are used to nowcast series, that is to predict low frequency macroeconomic releases (say end-of-quarter) for the current period (or within-quarter). See, for example, Giannone et al. (2008) and more generally the comprehensive survey on nowcasting in Bańbura et al. (2013).

A second is the construction of macroeconomic indicators which serve as a reference for policy-makers and economists. Examples include the Chicago Fed National Activity Index for the US, a monthly index designed to gauge overall economic activity and related inflationary pressure (Evans et al., 2002; Stock and Watson, 1999), and Eurocoin for the euro area, a monthly indicator that assesses economic activity free from short-run fluctuations (Altissimo et al., 2010).

A third is monetary policy and the identification of impulse response functions. A famous example is Forni and Gambetti (2010) which considers various anomalous empirical findings, labelled "puzzles", in previously undertaken studies. For instance, Sims (1992) finds, using structural vector autoregression (SVAR) analysis, that, after a monetary contraction, prices increase. Further, Eichenbaum and Evans (1995) and Grilli and Roubini (1996) find that exchange rates react with too long a delay relative to predictions of mainstream economic theory. Forni and Gambetti (2010) uses factor methods, which enable the researcher to handle a large amount of information and therefore avoid an important limitation of SVAR models known as non-fundamentalness. The puzzles are thereby resolved in the sense that the empirical results of Forni and Gambetti (2010) concur with mainstream theory.

A fourth is risk management and/or portfolio optimisation. A good example is Fan et al. (2013) wherein the primary aim is estimation of the covariance

matrix of an observed high-dimensional time series (e.g. of financial returns). This is achieved by imposing a "low rank plus sparsity" assumption on the covariance matrix and then regularising only its sparse component. To this end, the afore-mentioned factor model decomposition into common and idiosyncratic components serves as an ideal tool since it embodies perfectly the conditional sparsity structure that the authors seek. Indeed, Fan et al. (2013) makes the link with factor models and deploys relevant methods to great effect in the first stage of development of its so-called "Principal Orthogonal complEment Thresholding (POET) estimator" (Fan et al., 2013, p.3).

We hope that the foregoing discussion delivers an adequate flavour not only of the wide-ranging applicability of factor models but also their importance in various empirical settings. The interested reader is once again referred to the afore-mentioned surveys since they contain much additional background information. Our focus, at this stage, turns to an additional source of our enthusiasm for factor models. That is, the remarkable theoretical representation results that have been established in the field. Before we discuss those, however, we clarify commonly-used terminology and address the contemporary taxonomy of main types of factor models.

1.1.3 Exact Static Models

The work of Spearman (1904) entails what may be classified in modern-day parlance as an exact (or strict) static factor model. Exact, since all co-movements between observables are modelled via the factors alone, and static, since the model is designed only to explain independent data. A typical formulation for say an r factor model might be

$$x_{it} = \sum_{j=1}^{r} \lambda_{ij} f_{jt} + \xi_{it},$$

for, where $f_{1t}, ..., f_{rt}$ are referred to as the r common factors, $\xi_{1t}, ..., \xi_{Nt}$ as the N idiosyncratic components (with $r \ll N$), and where the factors and idiosyncratic components are assumed mutually orthogonal. The structure is static since factors are loaded contemporaneously and there is no temporal correlation in either

the factors or the idiosyncratic component. It is exact since the covariance of the vector of idiosyncratic components $(\xi_{1t}, ..., \xi_{Nt})'$ at time t is a diagonal matrix.

Credit for development of estimation methods for such models by maximum likelihood (ML) is due to Jöreskog (1969) and Lawley and Maxwell (1962), and by principal components analysis (PCA) to Tipping and Bishop (1999). These contributions all represent very important milestones in the development of the methodology for exact static factor models (and beyond). Nevertheless, the practical value of these models at least for present-day macroeconomics and finance is limited due to the overly restrictive nature of the assumptions. A good reference for examples of their use in social statistics is Bartholomew et al. (2011).

1.1.4 Exact Dynamic Models

With specific reference to stationary time series data, it is Geweke (1977) that is attributed with outlining an exact dynamic factor model. Dynamics are typically incorporated by allowing factors, and possibly also idiosyncratic components, to follow autoregressive processes. Given their structure, these models may readily be cast in state space form, and under Gaussianity, they lend themselves to estimation via the Kalman filter and likelihood based methods – see, for instance, Harvey (1990). Important contributions in time domain methods for estimation of these models were made by Watson and Engle (1983) and Quah and Sargent (1993), both of which make use of the expectation-maximisation (EM) algorithm. Key frequency domain references for estimation of exact static factor models include Geweke (1977) itself and Sargent and Sims (1977).

To illustrate the state space formulation, let us consider a model in which an r-dimensional vector of factors $\mathbf{f}_t$ follows a vector autoregression of order 1 – or a VAR(1) – process as follows

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\xi}_t \\ \mathbf{f}_t &= \mathbf{\Phi}\mathbf{f}_{t-1} + \mathbf{u}_t, \end{aligned}$$

where $\mathbf{\Lambda}$ is an $N \times r$ matrix of loadings, $\boldsymbol{\xi}_t$ is an N-dimensional vector, $\mathbf{\Phi}$ is an

$r \times r$ matrix of autoregressive coefficients, and $\mathbf{u}_t$ is a vector-valued white noise. While this is clearly not the most general framework available, it constitutes a reasonable way to introduce dynamics in the sense that (i) a VAR(p) model, for $p > 1$, can always be re-expressed as a VAR(1) by stacking the lagged terms; and (ii) the state equation can readily be augmented to incorporate dynamics in the idiosyncratic component too. Indeed, Stock and Watson (1988) uses just such a model to construct a coincident economic indicator for the US economy.

Nevertheless, we emphasise at this point that the covariance matrix of $\boldsymbol{\xi}_t$ remains diagonal in this model and it therefore appears overly restrictive for modern applications. The reason is that it forces all co-movements amongst the cross-section to occur via the common factors alone.

1.1.5 Approximate Static Models

The next question is whether one could design models without the restriction that features unique to a given cross-sectional unit are entirely uncorrelated with each other. Of course, there is a clear problem of logic here, namely that we cannot allow "too much" correlation otherwise (as mentioned previously) any idiosyncracy loses its meaning. Nevertheless, provided we place appropriate bounds on the cross-sectional pervasiveness of idiosyncratic components, we can indeed develop the theory for approximate static factor models (and subsequently also consider extensions to the dynamic case).

If we were to provide an illustrative model specification here, it would be similar to that presented under the previous category except (i) there would be no state equation, and (ii) there would be no requirement for the covariance matrix of $\boldsymbol{\xi}_t$ to be diagonal. An example of how identification may be carried out is as follows. Let us define an N-dimensional vector $\chi_t = \boldsymbol{\Lambda}\mathbf{f}_t$ to denote the common component of $\mathbf{x}_t$. Suppose $\psi_{i\chi}^{(N)}$ denotes the i-th largest eigenvalue of the covariance matrix of χ_t and analogously $\psi_{i\xi}^{(N)}$ of ξ_t, where the superscripts simply emphasise dependence of these quantities on N. Then, we could assume that (i) $\lim_{N\to\infty} \psi_{r\chi}^{(N)} = \infty$; and (ii) there exists a finite positive constant M independent of N such that $\psi_{1\xi}^{(N)} < M$. While this is a very intuitive method for disentan-

gling that which is common from that which is not, it is not the only way one could define pervasiveness and lack thereof. See, for instance, Bai and Ng (2002) and Fan et al. (2013), each of which frames constraints in ways that are subtly different to the above.

As regards estimation, PCA is well-suited to the static approximate setting. This is because PCA, an algebraic exercise, and factor analysis, a modelling exercise, both decompose the covariance matrix of $\mathbf{x}_t$ into the sum of a reduced rank matrix and whatever else remains. Combining PCA and factor analysis just seems like the natural thing to do. Key studies include Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986) which proposed use of PCA for approximate static factor models, and Stock and Watson (2002a,b) and Bai and Ng (2002) which find consistency and establish $\min(\sqrt{N}, \sqrt{T})$ rates for loadings and factor estimators. Bai (2003) obtains asymptotic distributions.

1.1.6 Approximate Dynamic Models

The theory for a much less restrictive class of factor models is developed in Forni et al. (2000), Forni and Lippi (2001), Forni et al. (2004), Forni and Lippi (2011) and Hallin and Lippi (2013).[2] These models (i) incorporate relaxations of orthogonality constraints on the idiosyncratic components, thereby allowing for the unique features of observable time series to indeed be mildly correlated; and (ii) incorporate dynamics, making them extremely well-suited for time series data. The key supposition is that our factors $\mathbf{f}_t$ are driven by a full-rank q-dimensional innovation process say ν_t, where $q \leq r \ll N$ and . The vector ν_t is typically interpreted as representing a set of mutually orthogonal standardised unobserved primitive shocks whose effects are propagated dynamically through the economic system and thereby responsible for the bulk of co-movements in macroeconomic variables now and in the future. An illustrative model formulation in this setting

[2] An additional related reference is Forni et al. (2005) with uses a modified version of the model in a forecasting context.

yields a common component of the form

$$\chi_{it} = \sum_{j=1}^{q} \sum_{s=1}^{\infty} B_{ij,s} \nu_{j,t-s},$$

where the $B_{ij,s}$ terms, for $1 \leq i \leq N, 1 \leq j \leq q$, quantify how the j-th shock is loaded onto the i-th time series at the s-th lag. By admitting potentially an infinite set of lags as above, we harbour an extremely flexible model structure.

The identification constraints alluded to in the previous section may be extended into this setting, for example, by (i) ensuring divergence of the q-th largest eigenvalue of the $N \times N$ spectral density matrix of χ_t as $N \to \infty$ for all frequencies; and (ii) ensuring that the largest eigenvalue of the $N \times N$ spectral density matrix of ξ_t is bounded from above uniformly with respect to the frequency.

It is worth emphasising that the approximate dynamic factor model is extremely general. In stark contrast to the various formulations alluded to in the foregoing paragraphs, the factor model decomposition discussed under this category is not so much a statistical model as it is a canonical representation. Indeed, we refer to Hallin and Lippi (2013, Theorem 1) which establishes that any second-order stationary process has a representation identical to the approximate dynamic factor model specification referred to above, a remarkable result.[3] A clear implication is that the approximate dynamic factor model encompasses all other models discussed heretofore.

1.1.7 Approximate Dynamic Models - with Restrictions

Given the benefits of such a flexible model structure, one approach is for researchers is to adopt a "holy grail" mindset and devote efforts purely into developing techniques suited to the approximate dynamic factor model discussed above. For instance, since we can no longer use PCA (which is unable by itself to handle dynamics) to estimate the model, Forni et al. (2000) proposes moving

[3] All that is needed is a 'bounded complexity' assumption that rules out the strange case of an infinite number of common shocks. (Hallin et al., 2020, Footnote 9).

to the frequency domain and using dynamic PCA (Brillinger, 1981, Chapter 9) instead. The "holy grail" approach, however, is not without limitations.

First, any statistical techniques necessarily involve an increase in technical sophistication and so theoretical results in pursuit of an ideal may be hard to come by (if not altogether impossible). For instance, frequency domain estimation is itself a complex enough subject but were we to bring in additional objectives (e.g. changepoint detection, missing data, hierarchical structures, etc.), the problem could get very big, very fast.

Second, solutions found, however remarkable, may remain somewhat unsatisfactory from an empirical perspective. For example, as phrased in Forni et al. (2015), "[...] estimators [of Forni et al. (2000)] are two-sided filters whose performance at the end of the observation period or for forecasting purposes is rather poor. No such problem arises with estimators based on standard principal components".

Third, it is unlikely that any technique (no matter how noble the author's intentions) will truly refrain from imposing any implicit or explicit additional constraints over and above what is needed for the barebones canonical representation result. For example, dynamic PCA is based on a factorisation of spectral density matrices. It follows that even the seminal work of Forni et al. (2000) needs to add assumptions to ensure that all $\mathbf{x}_t$ processes, for any possible size of cross-section, admit spectral densities. In absence of this assumption, the estimators of Forni et al. (2000) cannot be operationalised; but no such assumption is needed for "Theorem 1" of Hallin and Lippi (2013). Nevertheless, as previously noted, the Forni et al. (2000) approach is typically referred to as the "general..." or "generalised dynamic factor model".

Of course, we do not give up disheartened. Rather, we do the best we can. What that means in practice is that we typically sacrifice some generality in pursuit of advancing the state of the art. One approach is to assume that the space spanned by the factors at any time t has a finite-dimension r as N tends to infinity. Under this restriction, Forni et al. (2009) establishes that one can always re-express a dynamic model, say where q shocks loaded with p lags, into static form akin to that used by Stock and Watson (2005) and Bai and Ng (2007). The approach is

very popular since it yields a a very tractable state space formulation in which the measurement equation consists of a static approximate factor model characterised by $r = q(p+1)$ factors and the state equation comprises a vector autoregression. We do not present illustrative equations at this stage since these will shortly be laid out in Section 2.1 of the following chapter. For now, we note simply the implication that several aforesaid techniques, e.g. PCA and Ordinary Least Squares (OLS), become available once again. The Forni et al. (2009) approach is sometimes referred to as the "restricted generalised dynamic factor model".

Its popularity notwithstanding, the Forni et al. (2009) approach is by no means a cure-all. The finite-dimension assumption can be restrictive since it rules out certain (even quite elementary) factor loading patterns, examples of which can be found in Forni and Lippi (2011, p.23) and Forni et al. (2015, p.360). Leveraging the work of Anderson and Deistler (2008) on singular stationary processes with a rational spectrum, the latter studies advocate a completely different approach. That is, they show that under alternative assumptions, *inter alia* that the common component has a reduced-rank spectral density that exists and is rational, one-sided estimators based on PCA in the frequency domain can be obtained even when the space spanned by the common component is infinite-dimensional. See also Forni et al. (2017) for a discussion of the asymptotics of estimators within this setting. The Forni et al. (2015, 2017) approach is sometimes referred to as the "unrestricted generalised dynamic factor model".

This brings to a close our review of the salient features of the various models in the literature on factor analysis for high-dimensional time series. We believe the foregoing discussion will serve as adequate background for the terminology and models to be encountered in the remainder of our work. We turn our attention at this point to providing a description of associated statistical methods.

1.2 Synopsis of the Literature on Methods

In this section we provide an overview of key statistical methods for factor models for high-dimensional time series. As mentioned earlier, there exist studies

in areas as wide-ranging as missing data, jagged-edge data, block structures, non-stationarity, local stationarity, bootstrapping, high-frequency data, etc. We confess that in a field so bountiful, there is truly too much to say and there are too many to cite. For this reason, we limit our scope below purely to those sub-areas most relevant to this thesis: estimation, changepoint-detection, and determination of the number of factors.

1.2.1 Estimation

We touched upon key estimators at various places in Section 1.1 and aim to avoid excessive repetition below. In general, though, estimation methods for approximate dynamic factor models (with restrictions) could arguably, and at least for the purpose of this thesis, be grouped into two broad categories. The first of these categories is based on PCA and OLS estimation of vector autoregressions. The second is based on quasi-maximum likelihood (QML) estimation. We survey these two categories now.

1.2.1.1 PCA-based methods – Time Domain

Much of the literature in this area originated in the static approximate factor model context. The initial focus on static models is understandable since PCA (which targets the covariance matrix alone) overlooks serial correlation (which is captured by the entire autocovariance sequence). Nevertheless, most asymptotic results, such as the $\min(\sqrt{N}, \sqrt{T})$ consistency rate, carry through to extensions which do allow for serial correlation in factors and/or idiosyncratic components. As a result, the studies mentioned in Section 1.1.5 serve as extremely important building blocks for the dynamic setting too.

The most popular way of incorporating dynamics (at least in the empirical literature) is through judicious use of the state space formulation espoused by Forni et al. (2009).[4] An extremely influential study is Doz et al. (2011) which proposes

[4]While other approaches based on eigen-analysis do exist, we do not dwell on them here. See, for example, Lam et al. (2011) and Lam and Yao (2012), which exploit information in

a two-step estimator, wherein the first step estimates the measurement equation via PCA and the state equation via OLS.

For instance, say we have a model

$$\begin{aligned}\mathbf{x}_t &= \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\xi}_t \\ \mathbf{f}_t &= \mathbf{\Phi}\mathbf{f}_{t-1} + \mathbf{u}_t,\end{aligned}$$

with suitable "mildness" assumptions on $\boldsymbol{\xi}_t$. One can estimate the loadings and the factors via the constrained optimisation problem

$$\begin{aligned}\min_{\Lambda,\, \mathbf{f}_t} \quad & \frac{1}{NT}\sum_{t=1}^{T}(\mathbf{x}_t - \Lambda\mathbf{f}_t)'(\mathbf{x}_t - \Lambda\mathbf{f}_t), \\ \text{s.t.} \quad & \frac{\Lambda'\Lambda}{N} = I_r.\end{aligned}$$

The need for a constraint occurs due to the rotational indeterminacy arising due to the latency of the factors, and the constraint on the loadings presented above is only one of several possible normalisations. Nevertheless, we do not dwell on this point here. Rather, we note simply that the solution to the given problem is simply $\widehat{\mathbf{f}}_t = N^{-1}\widehat{\Lambda}'\mathbf{x}_t$ where $\widehat{\Lambda}$ is the matrix of eigenvectors of the $N \times N$ sample covariance matrix $T^{-1}\sum_1^T \mathbf{x}_t\mathbf{x}_t'$ corresponding to its r largest eigenvalues. Given $\widehat{\mathbf{f}}_t$, one can estimate the factor VAR using OLS.

1.2.1.2 PCA-based methods – Frequency Domain

Significant advances have also been made in the frequency domain. The advantage of spectral methods is that our focus need not be limited to the static or restricted settings and thus get us closer in spirit to the "holy grail" alluded to earlier. Of course, the studies mentioned in Section 1.1.6 are forerunners of research in this area, and the basic idea is as follows.

autocovariance matrices at non-zero lags (instead of just the covariance matrix as in PCA). In these studies, common factors are specifically defined as being responsible for *dynamic* movements in the observable time series, and to this extent, they also capture any serially correlated features of the idiosyncratic components.

Suppose we have a sequence of $N \times N$ autocovariance matrices denoted by $\{\Gamma_{\mathbf{x}}(h), h \in \mathbb{Z}\}$. Even if the same-period correlation of two series is negligible, the correlation could be still high at leads and lags. Ignoring this information, with a blinkered focus on the covariance matrix $\Gamma_{\mathbf{x}}(0)$ alone, can lead to significant losses in explanatory power. Dynamic PCA (Brillinger, 1981, Chapter 9) as espoused *inter alia* by Forni et al. (2000) overcomes this shortcoming by relying on the spectral density,

$$G_{\mathbf{x}}(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_{\mathbf{x}}(h) e^{-i\omega h},$$

for $\omega \in [-\pi, \pi)$, which encapsulates information in the entire autocovariance sequence. Analogous to conventional PCA, dynamic PCA is based on an eigen-decomposition of a non-parametric estimator of $G_{\mathbf{x}}(\omega)$. To estimate the common component by this method, we retain the eigenvectors corresponding to the q largest eigenvalues, carry out an inverse Fourier transform in order to obtain the so-called first q dynamic principal components, and project the data on the first q dynamic principal components to recover consistently the common component of each series.

Before we move on, it is worth pausing briefly to note that non-parametric estimation of the spectral density is not without its fair share of complexity. Indeed, we recount here an interesting anecdote from Hallin et al. (2020). With reference to the pioneering work of Forni et al. (2000), Hallin et al. (2020, p.11) concedes "Actually, the consistency result there is based (Equation (5) on page 545) on a wrong interpretation of a statement by Brockwell and Davis (Remark 1 of Section 10.4, page 353 of the 1991 edition). This was discovered several years later only [with credit ascribed to Giovanni Motta]".

As regards what the complexities are, we will in this thesis return to the issues surrounding non-parametric spectral density estimation in Section 2.6 of the upcoming chapter. For now, we note simply that a landmark study in the literature on lag-window estimators is Wu and Zaffaroni (2018). In fact, the statement quoted above from Hallin et al. (2020) goes on to add that the discovery of the aforesaid error was direct motivation for the latter study.

We wrap up our anecdote by considering what is the contribution of Wu and Zaffaroni (2018). As summarised succinctly by Forni et al. (2017, p.75), "[Wu and Zaffaroni (2018) establishes] that lag-window estimators of spectra and cross-spectra, under quite general assumptions on the processes and the kernel, are consistent, as $T \to \infty$, uniformly with respect to the [frequency ω], with rate $T/(B_T \log B_T)$, where B_T is the size of the lag window". This result is referred to by Hallin et al. (2020, p.11) as "[...] the mathematical cornerstone in the asymptotic analysis of [Forni et al. (2017)]". Indeed, just like Forni et al. (2017), our Chapter 2 will also benefit from this seminal result.

Finally, we note the finding of an overall min $\left(\sqrt{N}, \sqrt{T/(B_T \log B_T)}\right)$ consistency rate for dynamic PCA in this setting (Forni et al., 2017).

1.2.1.3 Likelihood-based methods - Time Domain

The second category of estimators we consider here are likelihood-based or QML methods. Key studies in this context include Doz et al. (2012) and Barigozzi and Luciani (2022).

As explained by Doz et al. (2012, p.1014), "[...] maximum likelihood estimation is clearly more appealing than principal components not only because it may lead to efficiency gains, but also, most importantly, because it provides a framework for incorporating restrictions derived from economic theory in the model. [...] For these reasons, establishing the properties of maximum likelihood estimators for factor models in large panels of time series is a relevant task from both the theory and applied point of view."

Doz et al. (2012) relies on likelihood-based methods to estimate, in an exact setting, a state space formulation of a dynamic factor model. In the spirit of QML methods (White, 1982), the approach is to treat the exact factor model as a mis-specified approximating model for the approximate (or generalised) case.

As regards computational methodology, Doz et al. (2012) leverages previous work by Doz et al. (2011) wherein the second step (of the proposed two-step estimator) estimates factors via the Kalman smoother. In particular, Doz et al. (2012) makes

use of the EM algorithm of Dempster et al. (1977) whereby (i) initialisation takes place via PCA and OLS as per the afore-mentioned first step in Doz et al. (2011); (ii) the E step makes use of the Kalman smoother as per the second step in Doz et al. (2011); and (iii) the M step estimates parameters via maximisation of the expected log-likelihood (conditional on the output of the E step). The two steps iterate until the algorithm terminates upon triggering a suitable convergence criterion. Doz et al. (2012) establishes consistency-with-rates for the proposed procedure.

Barigozzi and Luciani (2022) undertakes a meticulous scrutiny of the asymptotic properties of the QML estimator implemented via the EM algorithm. It establishes a min $\left(\sqrt{N}, \sqrt{T}\right)$ consistency rate (as well as other very interesting results including asymptotic normality) for estimators of the common component.

1.2.1.4 Likelihood-based methods - Frequency Domain

We now turn to the frequency domain, and this is precisely where the literature starts to thin. To the best of our knowledge, the only study that considers the possibility of an EM algorithm in the frequency domain is Fiorentini et al. (2018). Our impression of a lacuna in the literature is bolstered also by a reading of Barigozzi and Luciani (2022, "Remark 4") as well as the reasonably recent survey by Doz and Fuleky (2020, p.62) wherein Fiorentini et al. (2018) is the only study to be mentioned in this context.

Fiorentini et al. (2018) serves as a launch pad for our work in Chapter 2 in the sense that we too consider a spectral EM algorithm, whereby the E step is implemented via the Wiener-Kolmogorov (WK) smoother and the M step via maximisation of the Whittle Likelihood. To avoid repetition, we postpone a summary of the work of Fiorentini et al. (2018) and any associated references to the main body of Chapter 2. For now, we note simply that Fiorentini et al. (2018) does not consider the asymptotics of its proposed estimators, and that is a task that we attempt to undertake.

1.2.2 Changepoint Detection

This is another avenue of research within the factor models landscape where research has mushroomed (particularly in the aftermath of the 2008 crisis). Once again, it would be well beyond the scope of this review to comprehensively cover this very interesting area. The reason is that structural instabilities may successfully be modelled in several different ways (e.g. Markov-switching, slowly-varying loadings, etc.) which are not all relevant for this thesis.

In fact, it may surprise the reader that there have also been studies advocating that standard estimators found in the literature are *consistent even in the presence of temporal instability*. A case in point is the frequently-cited study by Stock and Watson (2002a) in which it is argued (see "Section 3" therein) that the proposed PCA estimators (which we already discussed above) are robust to stochastic drift in the factor loadings provided this drift is not too large and not pervasive across the cross-section – that is, if the instability is small enough and idiosyncratic enough, it will simply be eliminated asymptotically, by averaging across series.

Let us turn to Bates et al. (2013) for an example of the argument. Consider the temporally unstable static approximate single-factor model with one abrupt changepoint κ,

$$\begin{aligned}
\mathbf{x}_t &= \Lambda_t f_t + \mathbf{e}_t \\
\Lambda_t &= \Lambda_0 + \xi_t, \quad \text{where} \\
\xi_t &= \begin{cases} 0, & \text{for } t = 1, ..., \kappa \\ \Delta, & \text{for } t = \kappa + 1, ..., T. \end{cases}
\end{aligned}$$

Above, $\Delta \in \mathbb{R}^N$ may be interpreted as a shift parameter. Let us also define Δ_i as the i-th element of vector Δ for $i = 1, ..., N$. With such a model, we can clearly treat the instability as just another additive error term which, under the right conditions, can be appropriately dealt with. Indeed, under the assumption that that $|\Delta_i| \leq M$ for some finite positive constant M that is independent of i for all $i = 1, ..., N$, i.e. the instability is "small enough", Bates et al. (2013) establishes

that the standard PCA estimator is

$$O_p\left(\max\left(\max\left(\frac{1}{\sqrt{N}},\frac{1}{\sqrt{T}}\right),\frac{B/\sqrt{N}}{\sqrt{N}}\right)\right),$$

where $B \in \{1,...,N\}$ represents the number of series that undergo a break. It follows that if we add the restriction that at most $O(N^{1/2})$ series undergo a break, i.e. the instability is "idiosyncratic enough", we recover the standard consistency rate from the literature. If the findings of Stock and Watson (2002a) and Bates et al. (2013) are to be accepted, the implication is a serious one; that is, there is no need to expend efforts undertaking research on temporal instabilities.

The obvious question is whether the type and magnitude of instability considered by the above studies adequately captures what we observe in real-world scenarios. The message from Bates et al. (2013, p.290) is "[...] that the principal components estimator [is] robust to empirically relevant degrees of temporal instability in the factor loadings, although the precise quantitative conclusions depend on the assumed type of structural instability [...]". This is where the debate arises.

Yamamoto (2016, p.81) disagrees:

> ...we find that a significant portion [around 80] of 132 U.S. macroeconomic time series have structural changes in their factor loadings. Although traditional principal components provide eight or more factors, there are significantly fewer nonspurious factors.

As does Breitung and Eickmeier (2011, p.71):

> ...in empirical applications parameters may change dramatically due to important economic events, such as the collapse of the Bretton Woods system, or changes in the monetary policy regime, such as the conduct of monetary policy in the 1980s in the US or the formation of the European Monetary Union (EMU). There may also be more gradual but nevertheless fundamental changes in economic structures that may have led to significant changes in the comovements of variables, such as those related to globalization and technological progress. The common factors may become more (less) important for some of the

variables and, therefore, the loading coefficients attached to the common factors are expected to become larger (smaller). If one is interested in estimating the common components or assessing the transmission of common shocks to specific variables, ignoring structural breaks may give misleading results.

Indeed, several authors choose to pursue research on structural breaks. Historically, key studies in the field include Breitung and Eickmeier (2011), Chen et al. (2014), Cheng et al. (2016), Corradi and Swanson (2014), Han and Inoue (2014), Yamamoto and Tanaka (2015), Baltagi et al. (2017), and Ma and Su (2018), Barigozzi et al. (2018a) to name a few. However, all these studies consider the offline setting. That is, their aim is an in-sample detection of breaks. Even though these are not as relevant as the online case for our Chapter 3, we still glean a simple but key insight from several of these studies. We describe this insight next.

Specifically, we learn that a model with a single abrupt change in loadings at a given date has a representation as a model with constant loadings but a larger set of factors. To see this, we consider a one-factor model with a structural break.

$$\mathbf{x_t} = \begin{cases} \boldsymbol{\lambda_1} f_t + \mathbf{e_t}, & t \leq \kappa \\ \boldsymbol{\lambda_2} f_t + \mathbf{e_t}, & t > \kappa \end{cases}$$

If we define

$$g_{1t} = \begin{cases} f_t, & t \leq \kappa \\ 0, & t > \kappa \end{cases} \quad \text{and} \quad g_{2t} = \begin{cases} 0, & t \leq \kappa \\ f_t, & t > \kappa \end{cases},$$

we obtain $\mathbf{x_t} = \boldsymbol{\lambda_1} g_{1t} + \boldsymbol{\lambda_2} g_{2t} + \mathbf{e_t}$, an equivalent stable model. This insight will also underpin the procedure proposed in Chapter 3.

In contrast to the offline setting, the online case concerns the identification of breaks immediately as new data become available to the researcher. There is already a long tradition of literature (which originated outside time series) in the univariate setting for both dependent and independent data. See, for instance, Tartakovsky et al. (2014) for an overview. As regards time series, and specifi-

cally in the context of high-dimensionality, it would appear that development of detection methods for use with factor models would be a natural line of research.

To the best of our knowledge, however, the only study in this area so far is Barigozzi and Trapani (2020). It proposes a test centred around the property that a setup with an abrupt change in loadings (in a static approximate setting) is indistinguishable from one where there is stability in loadings but an emergence of a new factor (and simultaneous disappearance of an old). More formally, in an r factor model, a sudden change in loadings would result in spiking behaviour by the $(r+1)$-th eigenvalue. This behaviour is exploited and a randomised testing procedure (using rolling windows) is proposed in order to detect changepoints on a real-time basis.

We provide a detailed review of the literature on sequential changepoint detection in Section 3.2 below.

1.2.3 Determining Number of Factors

This is arguably the very first step in many applications involving factor analysis. Key studies in this area are reviewed below.

The forerunner is Bai and Ng (2002), which considers the approximate static setting. It considers the cross-sectional average of the estimated variance of the idiosyncratic component which is clearly minimised when the number of factors is chosen to be equal to the size of the panel N. In order to avoid this sort of obvious overparametrisation, a penalty is introduced, and this gives rise to a model selection criterion function that can be minimised in order to get a consistent estimate of the number of factors. Alessi et al. (2010) proposes a refinement to the Bai and Ng (2002) procedure by introducing into the penalty function a new parameter in order to fine tune its penalising power. We also have Hallin and Liška (2007) which proposes a criterion analogous to Alessi et al. (2010) but in the frequency domain in order to the tackle the approximate dynamic setting.

We note briefly that while Hallin and Liška (2007) mirrors the time domain work

of Alessi et al. (2010) in the frequency domain, there was no other frequency domain analogue to Bai and Ng (2002). Indeed, prior to Hallin and Liška (2007), the main proposal that existed in the literature for the approximate dynamic setting was simply a heuristic inspection-by-eye method put forward by Forni et al. (2000).

Other approaches look at eigenvalues directly. For instance, Onatski (2009) develops a test based on eigenvalues of the estimated spectral density matrix of data. Its test statistic effectively measures the curvature of the frequency domain scree slope Cattell (1966) at the "kink" under the alternative hypothesis. When the alternative hypothesis is true, the curvature at the kink asymptotically goes to infinity. In contrast, under the null, the curvature at the kink has a non-degenerate limiting distribution that does not depend on the model parameters. Ahn and Horenstein (2013) also attempts to mathematically formalise a search for the aforesaid kink in the scree. It proposes an "Eigenvalue Ratio" (ER) estimator, which is obtained by maximising the ratio of two adjacent eigenvalues arranged in descending order. Similarly, Onatski (2010) also proposed an estimator, named the "Edge Distribution" (ED) estimator, which estimates the number of factors using differenced eigenvalues. Both studies rely on interesting results from random matrix theory. Ahn and Horenstein (2013) and Onatski (2010) are based in the time domain and Onatski (2009) in the frequency domain.

The final study we mention to close our review of this area is Trapani (2018), which proposes a randomised sequential procedure to estimate the number of factors in a static approximate factor model. The work of Trapani (2018) is based in the time domain. Of course, we return to this study in further detail in Chapter 4 of this thesis. For now, we note simply that (to the best of our knowledge) there is no study that mirrors the work of Trapani (2018) in the frequency domain.

"In 1666, when Newton employed a prism to cast a rainbow on the wall..."
– Brillinger (1993)

Chapter 2

Estimation via a Spectral EM Algorithm

We develop a method to estimate an approximate dynamic factor model in which dynamics are exhibited in two ways: (i) factors follow an autoregression; and (ii) factors are loaded with a lag in the measurement equation. Within this framework, we propose a spectral EM algorithm, whereby we derive the E and M step equations in the frequency-domain. Following Fiorentini et al. (2018), our E step relies on the Wiener-Kolmogorov smoother, the frequency-domain counterpart of the Kalman smoother, and our M step is based on maximisation of the Whittle Likelihood with respect to the parameters of the model. Having outlined the E and M steps, we discuss initialisation of our procedure using dynamic PCA as per Forni et al. (2000). By leveraging results on lag-window estimators of spectral density by Wu and Zaffaroni (2018), we find consistency of our estimator of the common component as $N, T \to \infty$. We find rates commensurate with the literature, e.g. Forni et al. (2017), of $\min\{\sqrt{N}, \sqrt{B_T}, \sqrt{T/(B_T \log B_T)}\}$, where B_T is the size of the lag-window. Finally, we conduct simulations to find that our procedure performs as expected. We believe that by establishing a frequency-domain analogue to the EM algorithm of Doz et al. (2012) and Barigozzi and Luciani (2022), our study helps advance the field on QML methods for estimation of approximate dynamic factor models.

2.1 Introduction

We have witnessed a widespread surge in availability of large datasets in recent decades. Indeed, the catchphrase "big data" has firmly established its place in our lexicon. In macroeconometrics, for instance, we often encounter datasets involving high-dimensional time series; that is, with a large number of data points (T) on a large number of observable variables (N).

Dynamic factor models have garnered a lot of interest in this context. See, for instance, the survey in Stock and Watson (2016), and more recently, Doz and Fuleky (2020), as well as the references in each of these studies. A primary reason for the popularity of dynamic factor models is that they allow us to account for co-movements over time between large numbers of observable time series very parsimoniously. That is, through a small number (r) of latent factors thought to account for commonalities between the N time series. Any remaining movements that are unique to an individual series are modelled by way of an idiosyncratic component. The precise specifications of models deployed in practice vary in sophistication but, by holding $r \ll N$, all specifications seek to exploit the same central feature of dynamic factor models – a reduction in dimensionality. This gives empiricists a very convenient low-dimensional characterisation of economy-wide fluctuations, one that economic analysts and policy-makers can subsequently focus on.

Our study concerns estimation of the latent factors, common components, and parameters of dynamic factor models.[1] Specifically, we consider a model in which an N-dimensional vector of zero-mean weakly stationary time series $\mathbf{x}_t$ is driven by an r-dimensional vector of factors $\mathbf{f}_t$ and N-dimensional idiosyncratic component $\boldsymbol{\xi}_t$ as follows

$$\mathbf{x}_t = C(L)\mathbf{f}_t + \boldsymbol{\xi}_t,$$

[1] We note that we are not concerned, however, with estimation of exactly how many latent factors there are, which we assume throughout to be *a priori* known. Studies on estimation of the number of factors by themselves constitute an important yet distinct branch of the literature. See, for instance, Bai and Ng (2002), Hallin and Liška (2007), Onatski (2009), Alessi et al. (2010), Lam and Yao (2012), Ahn and Horenstein (2013), and Trapani (2018).

for $t = 1, ..., T$. In this so-called measurement equation, $C(L)$ is a one-sided $N \times r$ lag polynomial that allows lagged $\mathbf{f}_t$ to impact $\mathbf{x}_t$. This is one source of dynamics. Further, we postulate an autoregressive law of motion for the factors via $A(L)$, a one-sided $r \times r$ lag polynomial. Our second equation is

$$A(L)\mathbf{f}_t = \mathbf{u}_t$$

for $t = 1, ..., T$, and for some r-dimensional vector of innovations $\mathbf{u}_t$. This constitutes a second source of dynamics. The idiosyncratic ξ_{it}'s are permitted to be mildly cross-sectionally correlated and so the model has an approximate rather than an exact factor structure.[2]

A model such as ours is what is referred to in Stock and Watson (2016) as a dynamic factor model in *dynamic form*, that is, with reference to the fact that factors are loaded with lags in the measurement equation. Under the assumption that $C(L)$ is of finite degree, it is also possible to re-express the model in *static form*. That is, one could stack the lags of dynamically loaded factors $\mathbf{f}_t$ to form a broader set (with cardinality no smaller than r) of statically loaded factors, say $\mathbf{g}_t$, appearing contemporaneously in the measurement equation.[3]

As an example, suppose we have a single factor f_t loaded with a single lag along with an autoregression for the factors. Then, we may re-express the dynamic form of the model

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\lambda_0} f_t + \boldsymbol{\lambda_1} f_{t-1} + \boldsymbol{\xi}_t, \text{ and} \\ f_t &= \alpha_1 f_{t-1} + u_t, \end{aligned}$$

in static form by defining $\mathbf{g}_t = (f_t, f_{t-1})'$, $\Lambda = (\boldsymbol{\lambda_0} \ \boldsymbol{\lambda_1})$, $A = \begin{bmatrix} \alpha_1 & 0 \\ 1 & 0 \end{bmatrix}$, and

[2] See Assumption A6 for our specific characterisation of mildness.

[3] The finite degree condition on $C(L)$ implies that there exists a $\mathbf{g}_t$ with finite dimension.

$\boldsymbol{\eta}_t = \begin{pmatrix} 1 & 0 \end{pmatrix}' u_t$, whereby it follows that

$$\mathbf{x}_t = \Lambda \mathbf{g}_t + \boldsymbol{\xi}_t = \begin{bmatrix} \boldsymbol{\lambda_0} & \boldsymbol{\lambda_1} \end{bmatrix} \begin{bmatrix} f_t \\ f_{t-1} \end{bmatrix} + \boldsymbol{\xi}_t, \text{ and}$$

$$\mathbf{g}_t = A\mathbf{g}_{t-1} + \boldsymbol{\eta_t} = \begin{bmatrix} \alpha_1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} f_{t-1} \\ f_{t-2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_t,$$

for $t = 1, ..., T$. Their equivalence notwithstanding, these two distinct forms of dynamic factor models are typically estimated using distinct estimation methodologies. Indeed, the static form is more suited to estimation in the time-domain, and the dynamic form to estimation in the frequency-domain.

In the time-domain, the most popular estimation methodologies for approximate dynamic factor models can potentially be grouped into two broad categories. The first of these is based on PCA combined with OLS estimation of vector autoregressions (see, for example, Bai and Ng, 2006, 2007; Forni et al., 2009). The second is based on QML estimation typically implemented via the EM algorithm jointly with the Kalman smoother (see, for example, Barigozzi and Luciani, 2022; Doz et al., 2012). Both categories have been extensively studied. One of the key results in the time-domain literature assuming the existence of a static form of the model is the finding of $\min(\sqrt{N}, \sqrt{T})$ consistency for the estimator of the common component. See Bai (2003, Theorem 3) for PCA and Barigozzi and Luciani (2022, Theorem 1) for QML.

As regards the frequency-domain, the first of the above categories has certainly been well-studied. Indeed, conventional or static PCA, which approximates the sample covariance of $\mathbf{x}_t$ with a matrix of reduced rank, was famously generalised to the frequency domain by Brillinger (1981, Chapter 9). The procedure referred to as dynamic PCA approximates the estimated spectrum of $\mathbf{x}_t$ with a matrix of reduced rank. The method was popularised for use with approximate dynamic factor models with finite-dimensional factor spaces by Favero et al. (2005); Forni et al. (2000, 2004, 2005) and for infinite-dimensional factor spaces by Forni et al. (2017).

Clearly, the option to use dynamic PCA depends on first having a consistent

estimator of the spectrum. For this reason, it is worth nothing also at this stage the seminal result of Wu and Zaffaroni (2018) which establishes that lag-window estimators of spectra, under quite general restrictions, are consistent as $T \to \infty$, uniformly with respect to the frequency, with rate $\min(\sqrt{T/(B_T \log B_T)})$, where B_T is the size of the lag-window. The latter result is leveraged for use with approximate dynamic factor models with infinite-dimensional factor spaces by Forni et al. (2017) which finds an $(N, T \to \infty)$-consistency rate for its estimator of the common component of $\min(\sqrt{N}, \sqrt{T/(B_T \log B_T)})$ (see Equation 1.7 therein).

We turn once again to the second category of estimators and the question of whether one could develop a frequency-domain procedure for QML via the EM algorithm analogous to the time-domain counterpart considered in Doz et al. (2012) and Barigozzi and Luciani (2022). To the best of our knowledge, the only paper to consider such a procedure thus far is Fiorentini et al. (2018), which outlines a spectral EM algorithm. Following the latter, we too consider a spectral EM algorithm whereby (i) in the E step, factors are estimated by means of the Weiner-Kolmogorov (WK) smoother, which is the frequency-domain counterpart of the Kalman smoother (Hannan, 1970, Chapter III.7); and (ii) in the M step, parameters are estimated by maximisation of the Whittle frequency-domain approximation of the likelihood (Geweke and Singleton, 1981; Sargent and Sims, 1977).

As such, Fiorentini et al. (2018) provides an excellent starting point for our study. The key focus of that paper, however, is different to ours. Indeed, the main aim of Fiorentini et al. (2018) is to address estimation of models in which factors follow autoregressive moving-average (ARMA) processes as opposed to pure autoregressive (AR) processes as used in much of the literature (see, for example, Doz et al., 2012). To this end, the authors outline a spectral EM algorithm for an exact dynamic factor model with a single common factor following an ARMA process, leaving approximate dynamic factor models with multiple factors for future research. After the authors lay out the algorithm, their focus shifts to speeding up computation by developing an iterated indirect inference procedure based on a sequence of auxiliary OLS regressions. Finally, the authors conduct an empirical study of co-movements in US sectoral employment indicators following Quah and

Sargent (1993). Notably, the authors do not develop any asymptotic theory for their proposed procedure.

Our study aims to extend the groundwork laid out by Fiorentini et al. (2018) by developing our own spectral EM algorithm and establishing its asymptotic properties as $N, T \to \infty$. Unlike Fiorentini et al. (2018), however, we do not consider factor processes with moving-average (MA) components for the moment. Instead, we consider the dynamic form of the (restricted) approximate dynamic r-factor model as alluded to earlier in this introduction. Given this setup, we pursue the following avenues of research:

(i) Assuming parameters are known, we outline the E step. We then establish $\sqrt{N}$-consistency for factors estimated via the WK smoother pointwise in t for $t = 1, ..., T$ (see Theorem 2.1 below);

(ii) Assuming factors are known, we outline the M step. We then establish $\sqrt{T}$-consistency for loadings estimators estimated via maximisation of the Whittle likelihood for any given N (see Proposition 2.4 below);

(iii) We propose the use of dynamic PCA for initialisation of our procedure and discuss practical concerns behind non-parametric estimation of the spectral density matrix using lag-window estimators. We leverage the result from Wu and Zaffaroni (2018) to find a consistency rate of

$$\min(\sqrt{N}, \sqrt{B_T^{2\kappa}}, \sqrt{T/(B_T \log B_T)})$$

as $N, T \to \infty$, where B_T is the size of the lag window and κ is a parameter[4] summarising the smoothness of the kernel used (see Proposition 2.5 below);

(iv) We combine all our findings to obtain an overall consistency rate of

$$\min(\sqrt{N}, \sqrt{B_T}, \sqrt{T/(B_T \log B_T)})$$

as $N, T \to \infty$ for our spectral EM algorithm estimator of the common component (see Proposition 2.9 below).

[4]It is referred to as the Parzen exponent. To fix ideas, we note that κ is simply equal to 1 for the very commonly-used Bartlett kernel.

(v) Finally, we conduct a Monte Carlo study to numerically validate our findings and to also compare our common component estimator with other popular time- and frequency-domain alternatives. We find that our procedure performs very much as expected (see Section 2.9 below).

The remainder of this study is organised as follows. In Sections 2.2 and 2.3, we lay out much of our modelling structure and notation. In Section 2.4, we work towards building an E step via the WK smoother, and in Section 2.5, an M step via the Whittle likelihood. In Section 2.6, we consider initialisation via dynamic PCA. In Section 2.7, we outline the detailed steps for implementation of a spectral EM procedure. In Section 2.8, we summarise asymptotic properties. In Section 2.9, we present a numerical illustration of the procedure and conclude with a Monte Carlo study.

Our assumptions are laid out throughout the study as and when they are needed, and new notation is clarified as and when it appears. Proofs are contained in the end-of-chapter appendix.

2.2 Data-Generating Process

We begin with a detailed description of our assumptions on the data-generating process. We model X, an $N \times T$ rectangular array of observations, as follows.

Specification

Assumption A1. *There exists a real-valued double-indexed stochastic process of the form* $\Xi = \{x_{it} \in L_2(\Omega, \mathcal{A}, \mathcal{P}) | i \in \mathbb{N}, t \in \mathbb{Z}\}$ *where* i *denotes the cross-sectional index and* t *the temporal index and* $(\Omega, \mathcal{A}, \mathcal{P})$ *is a given probability space. For any* $\{\mathbf{x}_t = (x_{1t}, x_{2t}, ..., x_{Nt})' | t \in \mathbb{Z}\}, N \in \mathbb{N}$, *an* N*-dimensional sub-process of* Ξ*, we assume* $\mathbf{x}_t$ *is zero mean with finite (and non-degenerate) second-order moments* $\Gamma_{\mathbf{x}}(h) = E[\mathbf{x}_t \mathbf{x}'_{t-h}], h \in \mathbb{Z}$. *We assume* X *is a finite realisation of* Ξ.

We impose a decomposition of the form

$$\mathbf{x}_t = \boldsymbol{\chi}_t + \boldsymbol{\zeta}_t, \quad t \in \mathbb{Z}$$

where our observables are expressed as the sum of two unobserved unique and mutually orthogonal components, a *common* component $\boldsymbol{\chi}_t$ and an *idiosyncratic* component $\boldsymbol{\zeta}_t$. In particular, we have the following assumption.

Assumption A2. *For any $N \in \mathbb{N}$, there exist*

(a) a nested[5] *N-dimensional zero mean process $\boldsymbol{\phi}_t = (\phi_{1t}, \phi_{2t}, ..., \phi_{Nt})', t \in \mathbb{Z}$;*

(b) an r-dimensional zero mean process $\boldsymbol{\psi}_t = (\psi_{1t}, \psi_{2t}, ..., \psi_{rt})', t \in \mathbb{Z}$, for some finite positive integer r;

(c) an r-dimensional process $\mathbf{f}_t = (f_{1t}, ..., f_{rt})', t \in \mathbb{Z}$;

(d) filters $c_{i,\cdot k}(L) = c_{i,0k} + c_{i,1k}L$ with finite coefficients, and a nested $N \times r$ matrix polynomial, $C(L)$, whose $(i,k)^{th}$ entry is $c_{i,\cdot k}(L)$, $i = 1, ...N, k = 1, ..., r$;

(e) filters $\beta_{is}(L)$ with coefficients $\beta_{is,h}$, where $\sum_{s=1}^{\infty} \sum_{h=0}^{\infty} \beta_{is,h}^2 < \infty$ for all $i \in \mathbb{N}$, and a nested $N \times N$ matrix polynomial $B(L)$ whose $(i,s)^{th}$ entry is $\beta_{is}(L)$, $i, s = 1, ...N, h = 0, 1, ..., \infty$;

(f) finite coefficients a_{kl}, and an $r \times r$ matrix polynomial $A(L)$ whose $(k,l)^{th}$ entry is given by the $(k,l)^{th}$ entry of I_r minus $a_{kl}L$, $k, l = 1, .., r$;

(g) an $r \times r$ matrix U;

such that

(i) the vector $(\boldsymbol{\psi}_t', \boldsymbol{\phi}_t')', t \in \mathbb{Z}$ is independent and identically distributed (iid) and orthonormal; in particular, $\text{var}(\psi_{kt}) = \text{var}(\phi_{it}) = 1$, and $\text{cov}(\psi_{kt}, \phi_{i(t-h)}) = 0$, for $k = 1, .., r, i = 1, ...N, h \in \mathbb{Z}$;

(ii) all solutions of $\det(A(z)) = 0, z \in \mathbb{C}$, lie outside the unit ball;

(iii) $UE[\boldsymbol{\psi}_t \boldsymbol{\psi}_t']U'$ is some finite positive definite $r \times r$ matrix; and

[5] We use the word "nested" with reference to an increase in N.

(iv) for $t \in \mathbb{Z}$,

$$\boldsymbol{\chi}_t = C(L)\mathbf{f}_t;$$

$$A(L)\mathbf{f}_t = U\boldsymbol{\psi}_t;$$

$$\boldsymbol{\zeta}_t = B(L)\boldsymbol{\phi}_t.$$

We thus have $\mathbf{x}_t = C(L)\mathbf{f}_t + \boldsymbol{\zeta}_t$, an *approximate* dynamic r-factor model, where the dynamic factors $\mathbf{f}_t$ are loaded contemporaneously as well as with a single lag, and $\boldsymbol{\zeta}_t$ is allowed to be cross-sectionally and serially correlated. Under Assumptions A1 and A2, $\mathbf{x}_t$ is stationary. Denoting $\boldsymbol{\upsilon}_t = U\boldsymbol{\psi}_t$, we also have that the dynamic factors are zero mean and stationary with finite (and non-degenerate) second-order moments and follow a VAR(1) structure $A(L)\mathbf{f}_t = \boldsymbol{\upsilon}_t$.

Identification Restrictions

To disentangle the common and idiosyncratic components (as $N \to \infty$), we impose conditions characterising their pervasiveness and non-pervasiveness respectively.

Let $G_{\boldsymbol{\chi}}(\omega)$ and $G_{\boldsymbol{\zeta}}(\omega)$, for almost all $\omega \in [0, 1)$, be the $N \times N$ spectral density matrices of $\boldsymbol{\chi}_t$ and $\boldsymbol{\zeta}_t$.[6] Let $\lambda_i\{G_{\boldsymbol{\chi}}(\omega)\}$ and $\lambda_i\{G_{\boldsymbol{\zeta}}(\omega)\}$ denote their i^{th} largest eigenvalues, $i = 1, ..., N$. We refer to these as *dynamic eigenvalues* and note that they are real (and positive) since spectral density matrices are Hermitian.

One approach would be, in the spirit of Forni et al. (2000), to assume that $\lambda_r\{G_{\boldsymbol{\chi}}(\omega)\}/N$ is bounded from above and below uniformly with respect to the frequency as $N \to \infty$. Further, that $\lambda_1\{G_{\boldsymbol{\zeta}}(\omega)\}$ is bounded from above uniformly with respect to the frequency for any N. This approach permits us indeed to disentangle $\boldsymbol{\chi}_t$ from $\boldsymbol{\zeta}_t$ (as $N \to \infty$). However, it remains unspecified whether the

[6] We use complex exponentials of the form $e^{i2\pi\omega} = \cos(2\pi\omega) + i\sin(2\pi\omega)$ where i (in this context) denotes the imaginary unit. Since our complex exponentials are normalised to have period 1, we only consider $\omega \in [0, 1)$. Further, spectral densities are defined up to a set of frequency values contained in a Borel set with Lebesgue measure zero. Strictly speaking, "almost all" and/or "almost everywhere" terminology should accompany statements involving spectral densities. The terminology has no practical implications, however, and following Barigozzi et al. (2018b, Footnote 4), we omit it in this study.

divergence of dynamic eigenvalues of $\boldsymbol{\chi}_t$ arises due to the asymptotic properties of the loadings or of the factors (as $N \to \infty$). If the goal were to identify the factors and loadings separately, one would require additional assumptions.

Thus, we impose conditions right at the outset on the transfer function associated with the filter $C(L)$. Let a star superscript denote a complex conjugate transpose. By restricting convergence of $C^*(e^{-i2\pi\omega})C(e^{-i2\pi\omega})/N$ as $N \to \infty$ to an $r \times r$ identity matrix for any $\omega \in [0, 1)$, we can guarantee divergence of the r largest dynamic eigenvalues of $\boldsymbol{\chi}_t$. We state the assumption and then show this property below.

Let the ordered first r complex orthogonal unit-modulus eigenvectors of $G_{\mathbf{f}}(\omega)$ be denoted by the $N \times r$ matrix $P_{\mathbf{f}}(\omega)$. We denote the corresponding $r \times r$ diagonal matrix of eigenvalues as $\Lambda_{\mathbf{f}}(\omega)$.

Assumption A3. *For $\omega \in [0, 1)$,*

(i) $G_{\mathbf{f}}(\omega)$ has distinct eigenvalues; and

(ii) it holds that

$$\lim_{N\to\infty} \sup_{\omega} \left\| C^*(e^{-i2\pi\omega})C(e^{-i2\pi\omega})/N - I_r \right\| = 0,$$

where I_r is the $r \times r$ identity matrix.

Below, we state a result that follows from the preceding assumptions. It will be useful for various proofs. The result confirms asymptotic divergence linearly with N, and separation of the r largest dynamic eigenvalues of the common component as $N \to \infty$.

Proposition 2.1. *Given Assumptions A1-A3, there exist, for $i = 1, ..., r$, finite constants $\underline{M}_i^{\boldsymbol{\chi}}$ and $\overline{M}_i^{\boldsymbol{\chi}}$ independent of ω such that*

$$\begin{aligned} 0 < \underline{M}_r^{\boldsymbol{\chi}} \leq \frac{\lambda_r\{G_{\boldsymbol{\chi}}(\omega)\}}{N} \leq \overline{M}_r^{\boldsymbol{\chi}} \\ < \underline{M}_{r-1}^{\boldsymbol{\chi}} \leq \frac{\lambda_{r-1}\{G_{\boldsymbol{\chi}}(\omega)\}}{N} \leq \overline{M}_{r-1}^{\boldsymbol{\chi}} < \\ ... \\ < \underline{M}_1^{\boldsymbol{\chi}} \leq \frac{\lambda_1\{G_{\boldsymbol{\chi}}(\omega)\}}{N} \leq \overline{M}_1^{\boldsymbol{\chi}}, \end{aligned}$$

for each $i = 1, ..., r$.

Proof. See Appendix 2.10.1. □

The advantage of our approach is that, in conjunction with Assumption A4 below, Assumption A3 permits identification of the dynamic factors and the loadings up to some matrix lag polynomial with an associated transfer function that is an $r \times r$ complex diagonal matrix with unit-modulus diagonal entries, say $Q(\omega)$. The time domain analogue is identification up to some diagonal matrix Q where the diagonal entries are ± 1. Further, by also enforcing in Assumption A5 that the first row of $C(e^{-i2\pi\omega})$ is positive and real, we can achieve full identification. The time domain analogue is to assume that the first row of loadings is positive. These identification assumptions are quite reasonable since the dynamic factors do not necessarily have any particular economic meaning by themselves.

Let us define the ordered first r complex orthogonal unit-modulus eigenvectors of $G_{\boldsymbol{\chi}}(\omega)$ as $P_{\boldsymbol{\chi}}(\omega)$, an $N \times r$ matrix. We denote the corresponding $r \times r$ eigenvalue matrix as $\Lambda_{\boldsymbol{\chi}}(\omega)$. Let us also define $\mathbf{z}^{\boldsymbol{\chi}}(\omega)$ as the discrete-time Fourier transform (DTFT) of $\boldsymbol{\chi}_t$. That is,

$$\mathbf{z}^{\boldsymbol{\chi}}(\omega) = \sum_{t=-\infty}^{\infty} \boldsymbol{\chi}_t e^{-i2\pi\omega t}, \quad \omega \in [0, 1).$$

Similarly, let $\mathbf{z}^{\mathbf{f}}(\omega)$ be the DTFT of $\mathbf{f}_t$.

Assumption A4. *For any $\omega \in [0,1)$, it holds that*

$$\mathbf{z}^{\mathbf{f}}(\omega) = N^{-1/2} P_{\boldsymbol{\chi}}^{*}(\omega) \mathbf{z}^{\boldsymbol{\chi}}(\omega).$$

Given Assumption A4, we have a fixed structure for $G_{\mathbf{f}}(\omega)$:

$$\begin{aligned} G_{\mathbf{f}}(\omega) &= N^{-1} P_{\boldsymbol{\chi}}^{*}(\omega) G_{\boldsymbol{\chi}}(\omega) P_{\boldsymbol{\chi}}(\omega) \\ &= N^{-1} P_{\boldsymbol{\chi}}^{*}(\omega) P_{\boldsymbol{\chi}}(\omega) \Lambda_{\boldsymbol{\chi}}(\omega) P_{\boldsymbol{\chi}}^{*}(\omega) P_{\boldsymbol{\chi}}(\omega) \\ &= N^{-1} \Lambda_{\boldsymbol{\chi}}(\omega). \end{aligned}$$

So $G_{\mathbf{f}}(\omega)$ is a diagonal matrix with positive and real entries on the diagonal and the entries are distinct by part (i) of Assumption A3.[7] Moreover, since

$$\begin{aligned} G_{\boldsymbol{\chi}}(\omega) &= C(e^{-i2\pi\omega}) G_{\mathbf{f}}(\omega) C^{*}(e^{-i2\pi\omega}) \\ &= C(e^{-i2\pi\omega}) N^{-1/2} \Lambda_{\boldsymbol{\chi}}(\omega) N^{-1/2} C^{*}(e^{-i2\pi\omega}) \\ &= P_{\boldsymbol{\chi}}(\omega) \Lambda_{\boldsymbol{\chi}}(\omega) P_{\boldsymbol{\chi}}^{*}(\omega), \end{aligned}$$

we have that $C(e^{-i2\pi\omega}) = N^{1/2} P_{\boldsymbol{\chi}}(\omega)$. This also means that for $\omega \in [0,1)$ and for any N,

$$\sup_{w} \| C^{*}(e^{-i2\pi\omega}) C(e^{-i2\pi\omega})/N - I_r \| = 0,$$

which satisfies but also strengthens part (ii) of Assumption A3. That is, the above property arises now *for any* N, which is achievable since we defined the factors and the loadings in a manner such that they change with N.

To summarise, we still have the properties of divergence (as $N \to \infty$) and separation of the r non-zero dynamic eigenvalues of $\boldsymbol{\chi}_t$. In addition, we are able to identify the DTFT of the dynamic factors up to a complex diagonal matrix with unit-modulus entries. The next assumption permits full identification of the DTFT of the dynamic factors.[8]

The final remaining source of indeterminacy is that dynamic eigenvectors are

[7] In fact, with Assumption A4 in place, we could have merely asked for the diagonal entries of $G_{\mathbf{f}}(\omega)$ to be distinct in part (i) of Assumption A3.

[8] See, for instance, Hörmann et al. (2015), which considers Dynamic Functional PCA.

only defined up to post-multiplication by a complex diagonal matrix with unit-modulus diagonal elements, $Q(\omega)$. Let $[C(e^{-i2\pi\omega})]_{1k}$ for $k = 1, ..., r$ denote the entries in the first row of $C(e^{-i2\pi\omega})$. We choose to fix $Q(\omega)$ as follows.

Assumption A5. *For any $\omega \in [0, 1)$, and for any $N \times r$ matrix of ordered dynamic eigenvectors of $\boldsymbol{\chi}_t$, say $\Pi_{\boldsymbol{\chi}}(\omega)$, there exist $r \times r$ complex diagonal matrices with unit-modulus diagonal entries $Q(\omega)$, which depend on N and $P_{\boldsymbol{\chi}}(\omega)$, such that*

(i) $C(e^{-i2\pi\omega}) = N^{1/2}\Pi_{\boldsymbol{\chi}}(\omega) = N^{1/2}P_{\boldsymbol{\chi}}(\omega)Q(\omega)$*; and*

(ii) $[C(e^{-i2\pi\omega})]_{1k} \in \mathbb{R}_{\geq 0}$ *for* $k = 1, ..., r$.

Given Assumption A5, we can find $\Pi_{\boldsymbol{\chi}}(\omega)$ from $P_{\boldsymbol{\chi}}(\omega)$ by choosing a diagonal matrix $Q(\omega)$, such that for any $\omega \in [0, 1)$,

$$\text{diag}\left(Q(\omega)\right) = \left(\frac{\overline{[P_{\boldsymbol{\chi}}(\omega)]_{11}}}{|[P_{\boldsymbol{\chi}}(\omega)]_{11}|}, ..., \frac{\overline{[P_{\boldsymbol{\chi}}(\omega)]_{1r}}}{|[P_{\boldsymbol{\chi}}(\omega)]_{1r}|} \right),$$

where the numerators are the complex conjugates and the denominators are the moduli of the entries of the first row of any general $N \times r$ dynamic eigenvector matrix $P_{\boldsymbol{\chi}}(\omega)$.[9] The numerators ensure that the first row of $N^{1/2}P_{\boldsymbol{\chi}}(\omega)Q(\omega)$ will be positive and real, while the denominators ensure that the diagonal entries of $Q(\omega)$ are all unit-modulus. Specifically, the transformation is such that post-multiplication by $Q(\omega)$ replaces each entry in the first row of $P_{\boldsymbol{\chi}}(\omega)$ with its own modulus.

This completes our discussion of the common component. It remains now to ensure non-pervasiveness of the idiosyncratic component.

Assumption A6. *There exists a finite positive integer $M^{\boldsymbol{\zeta}}$ such that for $\omega \in [0, 1)$, it holds that $0 < \sup_\omega \lambda_1\{G_{\boldsymbol{\zeta}}(\omega)\} \leq M^{\boldsymbol{\zeta}}$ for all $N \in \mathbb{N}$.*

Let us define $g_{\boldsymbol{\zeta},is}(\omega)$ as the $(i, s)^{th}$ element of $G_{\boldsymbol{\zeta}}(\omega)$. The bound on the largest dynamic eigenvalue of the idiosyncratic component in A6 means that the average

[9] In addition to the definition stated above, we also set to zero any element of $Q(\omega)$ for which the divisor is zero.

column sum of the spectral density of $\boldsymbol{\zeta}_t$ is bounded. Indeed,

$$N^{-1}\sum_{s=1}^{N}\sum_{i=1}^{N} g_{\boldsymbol{\zeta},is}(\omega) = N^{-1}\boldsymbol{\iota}' G_{\boldsymbol{\zeta}}(\omega)\boldsymbol{\iota} \leq \lambda_1\{G_{\boldsymbol{\zeta}}(\omega)\} \leq M^{\boldsymbol{\zeta}} < \infty$$

where $\boldsymbol{\iota}$ is an N-dimensional vector of ones. It is in this sense that we limit cross-sectional and serial correlation in $\boldsymbol{\zeta}_t$ to be "mild" and thereby characterise the idiosyncratic component to be non-pervasive.

Remark 1. Approximating Model. Assumptions A1-A6 are sufficiently general to model cases of interest (say, just by allowing for a higher number of lags in the filter $C(L)$). See for example Stock and Watson (2005) and Bai and Wang (2015). However, we proceed with what we refer to as an "approximating" model, which is much simpler in respect of the stochastic behaviour of the idiosyncratic component. Our approach is similar, for instance, to that in Doz et al. (2011). The motivation is that an exact factor structure, ruling out any cross-sectional or serial correlation in the idiosyncratic component, greatly facilitates derivations of equations relating to our proposed estimators. Further, Gaussianity permits us to use likelihood-based methods. We make similar assumptions on the stochastic disturbance term in the factor equation. We will find consistency of our quasi-maximum likelihood estimators despite these forms of possible mis-specification. We outline our approximating model in detail below. △

2.3 The Approximating Model

We specify our *exact* dynamic r-factor model by a system of stochastic difference equations,

$$\begin{aligned}
\mathbf{x}_t &= \mathbf{c}_{01} f_{1,t} + \mathbf{c}_{11} f_{1,t-1} + \ldots + \mathbf{c}_{0r} f_{r,t} + \mathbf{c}_{1r} f_{r,t-1} + \boldsymbol{\xi}_t \\
f_{1,t} &= a_{11} f_{1,t-1} + \ldots + a_{1r} f_{r,t-1} + u_{1,t} \\
&\vdots \\
f_{r,t} &= a_{r1} f_{1,t-1} + \ldots + a_{rr} f_{r,t-1} + u_{r,t}
\end{aligned}$$

where

$\mathbf{x}_t$ is an $N \times 1$ vector of observables;

$f_{k,t}$ for every $k \in \{1, ..., r\}$ is a scalar unobserved common factor, and a vector of all r factors is denoted by $\mathbf{f}_t = (f_{1,t}, ..., f_{r,t})'$;

$\mathbf{c}_{0k}$ and $\mathbf{c}_{1k}$ are $N \times 1$ vectors of loadings corresponding to the k^{th} factor and the lagged k^{th} factor respectively; and

a_{kl} for every $k, l \in \{1, ..., r\}$ is a scalar coefficient on the l^{th} lagged factor in the autoregressive equation for the k^{th} factor.

Thus, our approximating model is identical to the one outlined in Assumption A2 with the notable exception of the two stochastic disturbance terms in the modelling equations. We assume a much simpler covariance structure for these terms as outlined in our assumption below. Further, we assume that the statement of Assumption A6 applies also to $\boldsymbol{\xi}_t$.

Assumption A7. *In the "approximating model" defined above,*

(i) $\boldsymbol{\xi}_t$ *is* $N \times 1$ *and* $\boldsymbol{\xi}_t \sim iid\ \mathcal{N}(0, \Gamma_{\boldsymbol{\xi}})$ *where* $\Gamma_{\boldsymbol{\xi}}$ *is a diagonal matrix; moreover,* $\boldsymbol{\xi}_t$ *is such that* $\text{diag}(\Gamma_{\boldsymbol{\xi}}) = \text{diag}(\Gamma_{\boldsymbol{\zeta}})$*;*

(ii) $u_{k,t} \sim \mathcal{N}(0, \gamma_k)$*, and* $u_{k,t}$ *is independent of* $u_{l,s}$ *for any* $k \neq l$ *or* $t \neq s$*. Alternatively, we have that* $\mathbf{u}_t = (u_{1,t}, ..., u_{r,t})' \sim iid\ \mathcal{N}(0, \Gamma_{\mathbf{u}})$ *for all* t *where* $\Gamma_{\mathbf{u}} = diag(\{\gamma_1, ..., \gamma_k\})$*; moreover,* $\mathbf{u}_t$ *is such that* $\text{diag}(\Gamma_{\mathbf{u}}) = \text{diag}(\Gamma_{\boldsymbol{\upsilon}})$*; and*

(iii) there exists a finite positive integer $M^{\boldsymbol{\xi}}$*, such that for* $\omega \in [0, 1)$*, it holds that* $0 < \sup_\omega \lambda_1\{G_{\boldsymbol{\xi}}(\omega)\} \leq M^{\boldsymbol{\xi}}$ *for all* $N \in \mathbb{N}$*.*

Concise Specification

We can also express the approximating model using lag polynomial notation. For $k = 1, ...r$, we have the filter

$$\mathbf{c}_k(L) = \mathbf{c}_{0k} + \mathbf{c}_{1k}L.$$

This is a lag polynomial of order 1 with dimension $N \times 1$.

Moreover, for $k, l = 1, ..., r$, we have the filter

$$a_{kl}(L) = a_{kl}L$$

which is a one-dimensional lag polynomial of order 1.

We can then construct the $N \times r$ matrix

$$C(L) = [\mathbf{c}_1(L) \quad ... \quad \mathbf{c}_r(L)]$$

and the $r \times r$ matrix

$$\begin{aligned} A(L) &= I_r - \begin{bmatrix} a_{11}(L) & ... & a_{1r}(L) \\ \vdots & \ddots & \vdots \\ a_{r1}(L) & ... & a_{rr}(L) \end{bmatrix} \\ &= \begin{bmatrix} 1 - a_{11}(L) & ... & -a_{1r}(L) \\ \vdots & \ddots & \vdots \\ -a_{r1}(L) & ... & 1 - a_{rr}(L) \end{bmatrix} \end{aligned}$$

so that our approximating model may be more concisely expressed as follows:

$$\mathbf{x}_t = C(L)\mathbf{f}_t + \boldsymbol{\xi}_t \tag{2.1}$$

$$A(L)\mathbf{f}_t = \mathbf{u}_t. \tag{2.2}$$

We gather all unknown parameters in the vector

$$\boldsymbol{\theta} = (\mathbf{c}'_{01}, ..., \mathbf{c}'_{0r}, \mathbf{c}'_{11}, ..., \mathbf{c}'_{1r}, a_{11}, ..., a_{rr}, \text{diag}(\Gamma_{\boldsymbol{\xi}})', \text{diag}(\Gamma_{\mathbf{u}})')'.$$

Spectral Densities

For a given h, we define $\Gamma_{\boldsymbol{\xi}}(h)$ and $\Gamma_{\mathbf{u}}(h)$, $h = -\infty, ..., 0, ..., \infty$, as the lag h autocovariance matrices of the stochastic disturbance terms in our approximating

model (where $\Gamma_{\boldsymbol{\xi}}(0) = \Gamma_{\boldsymbol{\xi}}$ and $\Gamma_{\mathbf{u}}(0) = \Gamma_{\mathbf{u}}$). The corresponding spectral density matrices are obtained as the DTFTs of the respective autocovariance matrices. Specifically, for any $\omega \in [0, 1)$,

$$G_{\boldsymbol{\xi}}(\omega) = \sum_{h=-\infty}^{\infty} \Gamma_{\boldsymbol{\xi}}(h) e^{-i2\pi\omega h}; \text{ and}$$
$$G_{\mathbf{u}}(\omega) = \sum_{h=-\infty}^{\infty} \Gamma_{\mathbf{u}}(h) e^{-i2\pi\omega h}.$$

Given the simplified structure of the stochastic disturbance terms in our approximating model, we have that for any $\omega \in [0, 1)$,

$$G_{\boldsymbol{\xi}}(\omega) = \Gamma_{\boldsymbol{\xi}}; \text{ and}$$
$$G_{\mathbf{u}}(\omega) = \Gamma_{\mathbf{u}}.$$

Discrete Fourier Transforms (DFTs)

Given X, our $N \times T$ rectangular array of observations, we define

$$\mathbf{z}_j^{\mathbf{x}} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{x}_t e^{-i2\pi\omega_j t}; \text{ and}$$
$$\mathbf{z}_j^{\mathbf{f}} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{f}_t e^{-i2\pi\omega_j t},$$

where the discretised frequency ω_j is defined as a grid of equally spaced values of $\omega \in [0, 1)$ specifically of the form $\omega_j = j/T$ for $j = 0, ..., T-1$. These are referred to as "Fourier" (or "fundamental") frequencies.

Then, by the linearity property of the DFT and the convolution theorem, we have that

$$\mathbf{z}_j^{\boldsymbol{\xi}} = \mathbf{z}_j^{\mathbf{x}} - C(e^{-i2\pi\omega_j})\mathbf{z}_j^{\mathbf{f}}; \text{ and}$$
$$\mathbf{z}_j^{\mathbf{u}} = A(e^{-i2\pi\omega_j})\mathbf{z}_j^{\mathbf{f}}$$

where the matrices containing the transfer functions associated with the linear filters are given by

$$C(e^{-i2\pi\omega_j}) = \begin{bmatrix} \mathbf{c}_1(e^{-i2\pi\omega_j}) & ... & \mathbf{c}_r(e^{-i2\pi\omega_j}) \end{bmatrix}$$

and

$$A(e^{-i2\pi\omega_j}) = \begin{bmatrix} 1 - a_{11}(e^{-i2\pi\omega_j}) & ... & -a_{1r}(e^{-i2\pi\omega_j}) \\ \vdots & \ddots & \vdots \\ -a_{r1}(e^{-i2\pi\omega_j}) & ... & 1 - a_{rr}(e^{-i2\pi\omega_j}) \end{bmatrix}$$

for any $\omega_j = j/T$ for $j = 0, ..., T-1$.

Filtrations

We define the information sets

$$\mathcal{D}_t = (\mathbf{x}_1, ..., \mathbf{x}_t), \text{ and}$$
$$\mathcal{F}_t = (\mathbf{x}_1, ..., \mathbf{x}_t, \mathbf{f}_1, ..., \mathbf{f}_t)$$

for any $t \in \{1, ..., T\}$.

2.4 Towards the E Step: WK Smoother (known parameters, fixed T)

In this section, we focus on developing the theoretical foundations for extraction of the factors (in preparation for our discussion of the E step). In particular, we present an expression for the conditional expectation of the DFT of the factors given (i) assumed values for the parameters $\boldsymbol{\theta}$, and (ii) the observed series. This is nothing but the WK smoother.[10] Once we extract the DFT of the factors in this way, we can easily return to the time domain using the inverse DFT. The

[10]Further, we present expressions for two additional conditional expectation terms involving the second-order structure of the factors.

key results of this section are Theorem 2.1 and its corollary which pertain to $\sqrt{N}$ consistency of the extracted signal assuming known parameters.

2.4.1 Computation of $\mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}}|\mathcal{D}_T]$

2.4.1.1 Wiener-Kolmogorov Smoother

Wiener-Kolmogorov theory of signal extraction indicates that optimal estimates of the latent factors are provided by conditional expectations that are formed given the observed data and characteristics of the models that are presumed to have generated them. A foundational reference is, of course, Wiener (1950). The WK smoother is nothing but the frequency-domain counterpart of the Kalman smoother (Hannan, 1970, Chapter III.7). A detailed summary of WK smoothing for multivariate time series with time-invariant state-space structure is Gómez (2007) which establishes equivalence between Wiener-Kolmogorov and Kalman methods within this setting. Finally, a specific application of the WK smoother to dynamic factor models similar to the one considered in our own study is Fiorentini et al. (2018).

For a given j, we obtain the conditional expectation of $\mathbf{z}_j^{\mathbf{f}}$ by the WK smoother as follows:

$$\mathbf{z}_j^{\mathbf{f}^W} = \mathsf{E}_{\boldsymbol{\theta}}\left[\mathbf{z}_j^{\mathbf{f}}|\mathcal{D}_T\right] = G_{\mathbf{f}}(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\mathbf{x}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}$$

where the $G_{\mathbf{x}}(\omega_j)$ is the spectral density of the observed series, and

$$G_{\mathbf{f}}(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\mathbf{x}}^{-1}(\omega_j)$$

is the transfer function of the WK smoother.

We note that the spectral density of the factors in our model is given by

$$G_{\mathbf{f}}(\omega_j) = [A^{-1}(e^{-i2\pi\omega_j})]\Gamma_{\mathbf{u}}[A^{-1}(e^{-i2\pi\omega_j})]^*$$

since we assume the factors follow a VAR(1) process $A(L)\mathbf{f}_t = \mathbf{u}_t$. This is in fact a diagonal matrix under Assumptions A1-A4.

2.4.1.2 Woodbury Formula

Given our approximating model, by the Woodbury formula, we have that

$$G_{\mathbf{x}}^{-1}(\omega_j) = G_{\boldsymbol{\xi}}^{-1}(\omega_j) - G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j),$$

where $W(\omega_j) = [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}$.

Thus,

$$\begin{aligned}
C^*(e^{-i2\pi\omega_j})G_{\mathbf{x}}^{-1}(\omega_j) &= C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j) - C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j}) \\
&\quad \times [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j) \\
&= \Big([G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})] - C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j}) \Big) \\
&\quad \times [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j) \\
&= G_{\mathbf{f}}^{-1}(\omega_j)W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)
\end{aligned}$$

so that the transfer function of the WK smoother may be written in terms of $W(\omega_j)$. That is,

$$G_{\mathbf{f}}(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\mathbf{x}}^{-1}(\omega_j) = W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)$$

Correspondingly, for a given j, the WK smoother becomes

$$\mathbf{z}_j^{\mathbf{f}^W} = W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}$$

where $W(\omega_j) = [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}$.

2.4.2 Computation of $\mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{f}^*}|\mathcal{D}_T]$

We note that, for a given j,

$$\mathsf{E}_{\boldsymbol{\theta}}\left[\mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{f}^*}|\mathcal{D}_T\right] = \mathbf{z}_j^{\mathbf{f}^W}\mathbf{z}_j^{\mathbf{f}^{W*}} + \mathsf{E}_{\theta}\left[\left(\mathbf{z}_j^{\mathbf{f}} - \mathbf{z}_j^{\mathbf{f}^W}\right)\left(\mathbf{z}_j^{\mathbf{f}} - \mathbf{z}_j^{\mathbf{f}^W}\right)^*|\mathcal{D}_T\right].$$

We consider the first term in the above expression. Based on the definition of the WK smoother, and noting the Hermitian nature of the relevant matrices, we compute the first term as follows:

$$\mathbf{z}_j^{\mathbf{f}^W}\mathbf{z}_j^{\mathbf{f}^{W*}} = W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}\mathbf{z}_j^{\mathbf{x}^*}G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})W(\omega_j).$$

As regards the second term, we use a textbook result on the conditional distribution of partitioned multivariate Gaussian random vectors. Simply stated, assuming standard regularity conditions, if random vector $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_{\mathbf{y}})$, where $\mathbf{y} \in \mathbb{R}^N$, its mean $\boldsymbol{\mu}$ and covariance Σ can be partitioned according to

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_a \\ \mathbf{y}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma'_{ab} & \Sigma_{bb} \end{pmatrix},$$

then $(\mathbf{y}_a|\mathbf{y}_b) \sim \mathcal{N}(\boldsymbol{\mu}_{a|b}, \Sigma_{a|b})$ where

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{y}_b - \boldsymbol{\mu}_b); \text{ and} \\ \Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma'_{ab}. \end{aligned}$$

$\Sigma_{a|b}$ is known as the Schur complement of matrix Σ with respect to block Σ_{bb}. If $\mathbf{y}_a \perp \mathbf{y}_b$, then $\Sigma_{a|b} = \Sigma_{aa}$.

We carry this result to the frequency domain, let $\mathbf{y}_a = (\mathbf{z}_j^{\mathbf{f}} - \mathbf{z}_j^{\mathbf{f}^W})$ and $\mathbf{y}_b = \mathbf{z}_j^{\mathbf{x}}$, and note that the cross spectral density of the factor estimation error and the observables is zero by design. Then, our second term is just the spectral density of the factor estimation error $(\mathbf{f}_t - \mathbf{f}_t^W)$, which we denote by $\Omega(\omega_j)$. That is,

$$\Omega(\omega_j) = \mathsf{E}_{\boldsymbol{\theta}}[(\mathbf{z}_j^{\mathbf{f}} - \mathbf{z}_j^{\mathbf{f}^W})(\mathbf{z}_j^{\mathbf{f}} - \mathbf{z}_j^{\mathbf{f}^W})^*].$$

Next, since the spectral density of the factors equals the sum of the spectral

densities of the smoothed factors and of the factor estimation error, we have

$$\begin{aligned}\Omega(\omega_j) &= \mathsf{E}_{\boldsymbol{\theta}}\left[\left(\mathbf{z}_j^{\mathbf{f}} - \mathbf{z}_j^{\mathbf{f}^W}\right)\left(\mathbf{z}_j^{\mathbf{f}} - \mathbf{z}_j^{\mathbf{f}^W}\right)^*\right] \\ &= G_{\mathbf{f}}(\omega_j) - G_{\mathbf{f}^W}(\omega_j) \\ &= G_{\mathbf{f}}(\omega_j) - G_{\mathbf{f}}(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\mathbf{x}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})G_{\mathbf{f}}(\omega_j).\end{aligned}$$

Using the fact that $G_{\mathbf{f}}(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\mathbf{x}}^{-1}(\omega_j) = W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)$, we see that

$$\begin{aligned}\Omega(\omega_j) &= G_{\mathbf{f}}(\omega_j) - W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})G_{\mathbf{f}}(\omega_j) \\ &= W(\omega_j)[W^{-1}(\omega_j)G_{\mathbf{f}}(\omega_j) - C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})G_{\mathbf{f}}(\omega_j)] \\ &= W(\omega_j)[[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]G_{\mathbf{f}}(\omega_j) \\ &\quad - C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})G_{\mathbf{f}}(\omega_j)] \\ &= W(\omega_j).\end{aligned}$$

Hence, our second term is $W(\omega_j)$. We treat it as just another parameter.

To summarise, for a given j, we use the formula

$$\mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{f}^*}|\mathcal{D}_T] = W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}\mathbf{z}_j^{\mathbf{x}^*}G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})W(\omega_j) + W(\omega_j).$$

2.4.3 Computation of $\mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{x}^*}|\mathcal{D}_T]$

We have, for a given j, that

$$\mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{x}^*}|\mathcal{D}_T] = \mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}}|\mathcal{D}_T]\mathbf{z}_j^{\mathbf{x}^*} = W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}\mathbf{z}_j^{\mathbf{x}^*}.$$

2.4.4 Population Results for WK Smoother

We begin this section with a remark that clarifies our meaning of uniform convergence with respect to frequencies.

Remark 2. UNIFORMITY. Say we have a sequence of r-dimensional vector-valued random variables depending on N, denoted $\mathbf{z}_j^{(N)}$, for $j = 0, ..., T-1$. This sequence is said to be $O_p(1)$ uniformly in j as $N \to \infty$, if for any $\epsilon > 0$, there exists a finite positive constant M_ϵ, independent of j, such that

$$\lim_{N\to\infty} \mathsf{Pr}\left(\left\|\mathbf{z}_j^{(N)}\right\| > \epsilon\right) < M_\epsilon$$

for any $j = 0, ..., T-1$. Indeed, since the constant M_ϵ is independent of j, we have specifically also that

$$\lim_{N\to\infty} \max_{0\le j\le T-1} \mathsf{Pr}\left(\left\|\mathbf{z}_j^{(N)}\right\| > \epsilon\right) < M_\epsilon.$$

This version of uniformity (which places a bound on the maximum of the probabilities) is weaker than the classical definition (in which the bound is on the probability of the maximum). We use this approach following Hallin and Liška (2007).

Analogously, we refer to a sequence $\mathbf{z}_j^{(N)}$ as being uniformly bounded in mean-square with respect to j if, as $N \to \infty$, there exists a finite positive constant M, independent of j, such that

$$\lim_{N\to\infty} \max_{0\le j\le T-1} \mathsf{E}\left(\left\|\mathbf{z}_j^{(N)}\right\|^2\right) < M,$$

or equivalently,

$$\max_{0\le j\le T-1} \mathsf{E}\left(\left\|\mathbf{z}_j^{(N)}\right\|^2\right) = O(1).$$

△

Proposition 2.2. *Under Assumptions A1-A7, and assuming parameters $\boldsymbol{\theta}$ are known, for any $T \in \mathbb{N}$, it holds that as $N \to \infty$,*

$$\max_{0\le j\le T-1} \mathsf{E}_{\boldsymbol{\theta}}\left(N\left\|\mathbf{z}_j^{\mathbf{f}^W} - \mathbf{z}_j^{\mathbf{f}}\right\|^2\right) = O\left(1\right).$$

Proof. See Appendix 2.10.2. □

For completeness, we also report below the implied result on convergence in probability. That is, the WK smoother is weakly consistent with rate $\sqrt{N}$ uniformly with respect to the frequencies.

Corollary. *Under Assumptions A1-A7, and assuming parameters $\boldsymbol{\theta}$ are known, for any $T \in \mathbb{N}$, it holds that as $N \to \infty$,*

$$\sqrt{N} \left\| \mathbf{z}_j^{\mathbf{f}^W} - \mathbf{z}_j^{\mathbf{f}} \right\| = O_p(1)$$

uniformly (as defined in Remark 2) with respect to j, where $j = 0, ..., T-1$.

Proof. The proof follows from an application of Chebyshev's inequality. □

It remains now to revert to the time domain. Given our frequency domain estimator $\mathbf{z}_j^{\mathbf{f}^W}$, we define our time domain estimator of the factors as

$$\mathbf{f}_t^W = \frac{1}{\sqrt{T}} \sum_{j=0}^{T-1} \mathbf{z}_j^{\mathbf{f}^W} e^{i2\pi\omega_j t}$$

for $t = 1, ..., T$, where $\omega_j = j/T$ for $j = 0, ..., T-1$.

Theorem 2.1. *Under Assumptions A1-A7, and assuming parameters $\boldsymbol{\theta}$ are known, for any $T \in \mathbb{N}$ and a given time period t, it holds that as $N \to \infty$,*

$$\mathsf{E}_{\boldsymbol{\theta}} \left(N \left\| \mathbf{f}_t^W - \mathbf{f}_t \right\|^2 \right) = O(1)$$

pointwise with respect to t, where $t = 1, ..., T$.

Proof. See Appendix 2.10.3. □

Corollary. *Under Assumptions A1-A7, and assuming parameters $\boldsymbol{\theta}$ are known,*

with some fixed $T \in \mathbb{N}$, it holds that as $N \to \infty$,

$$\sqrt{N} \left\| \mathbf{f}_t^W - \mathbf{f}_t \right\| = O_p(1)$$

pointwise with respect to t, where $t = 1, ..., T$.

Proof. The proof follows from an application of Chebyshev's inequality. □

In other words, using the WK smoother, we can recover the factors consistently with rate $\sqrt{N}$ pointwise in t, for $t = 1, ..., T$. This rate is identical to that obtained for equivalent time domain methods. For instance, Doz et al. (2011, Proposition 1) finds the same rate for the Kalman smoother.

Finally, we present for completeness some associated results on the second-order terms too.

Proposition 2.3. *Under Assumptions A1-A7, and assuming parameters $\boldsymbol{\theta}$ are known, for any $T \in \mathbb{N}$, it holds that as $N \to \infty$,*

(i) $\max_{0 \leq j \leq T-1} \mathsf{E}_{\boldsymbol{\theta}} \left(N \left\| \mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} | \mathcal{D}_T] - \mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \right\|^2 \right) = O(1)$*; and*

(ii) $\max_{0 \leq j \leq T-1} \mathsf{E}_{\boldsymbol{\theta}} \left(N \left\| \mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{x}^*} | \mathcal{D}_T] - \mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{x}^*} \right\|^2 \right) = O(1)$.

Proof. See Appendix 2.10.4. □

2.5 Towards the M Step: Likelihood (known factors, fixed N)

In this section, we focus on developing the theoretical foundations for estimation of the parameters (in preparation for our discussion of the M step). We begin by making a small aside in order to review the objective function for the standard EM algorithm following Dempster et al. (1977) and Wu (1983).

2.5.1 Objective Function for the EM Algorithm

Let us consider the (quasi-) maximum likelihood estimator of $\boldsymbol{\theta}$ given by

$$\arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(X;\boldsymbol{\theta}\right)$$

where $\mathcal{L}\left(X;\boldsymbol{\theta}\right)$ may be interpreted as the marginal likelihood of $\boldsymbol{\theta}$ given observations X that may be obtained from the complete likelihood $\mathcal{L}\left(\mathcal{F}_T;\boldsymbol{\theta}\right)$ of $\boldsymbol{\theta}$ given observations X *and* latent factors, say denoted by F. Of course, the marginal likelihood is unknown but, given the structure of the approximating model, the complete likelihood can indeed be analysed. Taking some liberties with notation (throughout this brief expository section), the following relation holds

$$\mathcal{L}\left(X;\boldsymbol{\theta}\right) = \int \mathcal{L}\left(X,F;\boldsymbol{\theta}\right) dF.$$

The interpretation of the integral is, of course, that we are interested in the likelihood of $\boldsymbol{\theta}$ given X across all possible values of F.

For any arbitrary density function of F, say $q(F)$, the right hand side of the above expression may be manipulated as follows,

$$\begin{aligned}\int \mathcal{L}\left(X,F;\boldsymbol{\theta}\right) dF &= \int \mathcal{L}\left(X,F;\boldsymbol{\theta}\right) \frac{q(F)}{q(F)} dF \\ &= \mathsf{E}\left[\frac{\mathcal{L}\left(X,F;\boldsymbol{\theta}\right)}{q(F)}\right],\end{aligned}$$

where the expectation is under the parameters governing $q(F)$.

Moreover, one typically does not maximise the likelihood; rather, one tends to focus on the log likelihood. So let us consider instead

$$\begin{aligned}\ell(X;\boldsymbol{\theta}) &= \log \mathcal{L}(X;\boldsymbol{\theta}) \\ &= \log \mathsf{E}\left[\frac{\mathcal{L}\left(X,F;\boldsymbol{\theta}\right)}{q(F)}\right] \\ &\geq \mathsf{E}\left[\log \frac{\mathcal{L}\left(X,F;\boldsymbol{\theta}\right)}{q(F)}\right],\end{aligned}$$

where the lower bound to $\ell(X;\boldsymbol{\theta})$ is found due to Jensen's inequality.

Since it is not possible to maximise $\ell(X;\boldsymbol{\theta})$ directly, the goal will be to maximise the log likelihood by repeatedly constructing a lower bound for it and then maximising that instead. It seems natural then that the choice of $q(F)$ should be guided by ensuring that the lower bound obtained above is tight and holds with equality. Of course, this will be the case if $\frac{\mathcal{L}(X,F;\boldsymbol{\theta})}{q(F)}$ is independent of F (and thus constant with respect to the expectation).

Indeed, we can set $q(F)$ to be the conditional distribution of F given X, $\mathcal{L}(F|X;\boldsymbol{\theta})$. Then, with an initial guess for parameters, say $\boldsymbol{\theta}^{(0)}$,

$$\begin{aligned}\ell(X;\boldsymbol{\theta}) &= \mathsf{E}_{\boldsymbol{\theta}^{(0)}}\left[\ell\left(X,F;\boldsymbol{\theta}\right)|X\right] - \mathsf{E}_{\boldsymbol{\theta}^{(0)}}\left[\ell\left(F|X;\boldsymbol{\theta}\right)|X\right] \\ &= \mathsf{E}_{\boldsymbol{\theta}^{(0)}}\left[\left(\ell\left(X|F;\boldsymbol{\theta}\right) + \ell\left(F;\boldsymbol{\theta}\right)\right)|X\right] - \mathsf{E}_{\boldsymbol{\theta}^{(0)}}\left[\ell\left(F|X;\boldsymbol{\theta}\right)|X\right] \\ &= \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(0)}) - \mathcal{H}(\boldsymbol{\theta};\boldsymbol{\theta}^{(0)}),\ \text{say,}\end{aligned}$$

where the "$\mathcal{Q}$-function" above is referred to as an auxiliary function. It is typically interpreted as a two-parameter family of curves enveloped by the likelihood function.

To help find $\arg\max_{\boldsymbol{\theta}} \ell\left(X;\boldsymbol{\theta}\right)$, we refer to Dempster et al. (1977), wherein the EM algorithm is outlined. The key requirement is that the likelihood is uniformly bounded from above for any $\boldsymbol{\theta}$ in the parameter space. Briefly, this is an iterative procedure with two steps to be repeated at each iteration of the algorithm. It is based on the idea that increasing the $\mathcal{Q}$-function increases the likelihood function.

Say we have $\boldsymbol{\theta}^{(k)}$, an estimated value of the parameters at a given iteration $k \geq 0$, we alternate between

E step: compute $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = \mathsf{E}_{\boldsymbol{\theta}^{(k)}}\left[\left(\ell\left(X|F;\boldsymbol{\theta}\right)|X\right] + \mathsf{E}_{\boldsymbol{\theta}^{(k)}}\left[\ell\left(F;\boldsymbol{\theta}\right)\right)|X\right]$; and

M step: compute $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$,

until a suitable convergence criterion has been satisfied.

Let us denote the final estimate of the parameters by $\boldsymbol{\theta}^{(EM)}$.

Of particular note for our purpose is that for any $k \geq 0$,

$$\ell\left(X;\boldsymbol{\theta}^{(k+1)}\right) - \ell\left(X;\boldsymbol{\theta}^{(k)}\right) = \mathcal{Q}(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}) - \left[\mathcal{H}(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}) - \mathcal{H}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\right],$$

where the first term is non-negative since $\boldsymbol{\theta}^{(k+1)}$ maximises $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ by construction. Further, the second term is no greater than zero as shown in Dempster et al. (1977, Lemma 1). The key advantage of the EM algorithm follows, namely that the log-likelihood is monotonically increasing in k. Indeed, given continuity of the $\mathcal{Q}$-function and if the $\mathcal{Q}$-function is not trapped at any point that is a stationary point but not a local maximum of the likelihood, the algorithm is *guaranteed* to converge to a local maximum as $k \to \infty$ for any N and T (Wu, 1983, Theorem 3).[11]

Thus, in this study, we focus only on the $\mathcal{Q}$-function with respect to parameter $\boldsymbol{\theta}$ and not also any additional terms in $\ell(X;\boldsymbol{\theta})$. Indeed, in the sections below, we focus on spelling out the ingredients for our own $\mathcal{Q}$-function. That is, we begin by considering a frequency domain approximation of $\ell(X,F;\boldsymbol{\theta}) = \ell\left(X|F;\boldsymbol{\theta}\right) + \ell\left(F;\boldsymbol{\theta}\right)$.

Of course, the concern that we attain a local rather than the global maximum remains outstanding. Indeed, as phrased in Wu (1983, p. 97), "[a]lthough a global maximization of Q is involved in the M step, the other term H in $L = Q - H$ may not cooperate". In practice, however, this will not be a concern for us, and we discuss the reason in Remark 3 below.

A general discussion pertaining to conditions for convergence guarantees for the EM algorithm, based chiefly on Dempster et al. (1977), Wu (1983) and Balakrishnan et al. (2017), is presented in Appendix 2.11. The latter expands upon several results that are alluded to above.

[11] For a unimodal concave and continuous likelihood which is bounded from above and has a single stationary point, the global maximum is attained.

2.5.2 Whittle Likelihood

We consider the Whittle likelihood which is formed from the distribution of the DFT of the data. Specifically, with Fourier frequencies ω_j for $j = 0, ..., T-1$, the Whittle spectral approximation to the complete log-likelihood function is

$$\begin{aligned}\ell(X, F; \boldsymbol{\theta}) \simeq & -\frac{1}{2}\sum_{j=0}^{T-1} \log |G_{\boldsymbol{\xi}}(\omega_j)| - \frac{1}{2}\sum_{j=0}^{T-1} \mathbf{z}_j^{\boldsymbol{\xi}^*} G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\xi}} \\ & -\frac{1}{2}\sum_{j=0}^{T-1} \log |G_{\mathbf{u}}(\omega_j)| - \frac{1}{2}\sum_{j=0}^{T-1} \mathbf{z}_j^{\mathbf{u}^*} G_{\mathbf{u}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{u}},\end{aligned}$$

where the we exclude the constant terms for brevity.[12]

Since we have assumed zero cross-sectional correlation in the idiosyncratic component, the autocovariance sequence of $\boldsymbol{\xi}_t$ is given by diagonal matrices.[13] Thus, $G_{\boldsymbol{\xi}}(\omega_j)$ is also diagonal. Hence, we can decompose the first line in $\ell(X, F; \boldsymbol{\theta})$ as the sum of N univariate components. Analagously, since there is zero correlation between common shocks, the autocovariance sequence of $\mathbf{u}_t$ is given by diagonal matrices. Thus, $G_{\mathbf{u}}(\omega_j)$ is diagonal. Hence, we can decompose the second line in $\ell(X, F; \boldsymbol{\theta})$ as the sum of r univariate components. We obtain

$$\begin{aligned}\ell(X, F; \boldsymbol{\theta}) \simeq & \sum_{i=1}^{N}\left[-\frac{1}{2}\sum_{j=0}^{T-1} \log |G_{\xi_i}(\omega_j)| - \frac{1}{2}\sum_{j=0}^{T-1} z_j^{\xi_i^*} G_{\xi_i}^{-1}(\omega_j) z_j^{\xi_i}\right] \\ & + \sum_{k=1}^{r}\left[-\frac{1}{2}\sum_{j=0}^{T-1} \log |G_{u_k}(\omega_j)| - \frac{1}{2}\sum_{j=0}^{T-1} z_j^{u_k^*} G_{u_k}^{-1}(\omega_j) z_j^{u_k}\right].\end{aligned}$$

Above, we use $G_{\xi_i}(\omega_j)$ to denote the i^{th} element on the diagonal of $G_{\boldsymbol{\xi}}(\omega_j)$, and $G_{u_k}(\omega_j)$ to denote the k^{th} element on the diagonal of $G_{\mathbf{u}}(\omega_j)$.

[12] See, for example, Krafty and Collinge (2013) and Fiorentini et al. (2018).

[13] This would be true even if we were to allow for serial correlation in the idiosyncratic component.

Plugging in the DFTs,

$$\begin{aligned}\ell(X, F; \boldsymbol{\theta}) \simeq \sum_{i=1}^{N} \Bigg[& -\frac{1}{2} \sum_{j=0}^{T-1} \log |G_{\xi_i}(\omega_j)| \\ & - \frac{1}{2} \sum_{j=0}^{T-1} [z_j^{x_i} - C_i(e^{-i2\pi\omega_j}) \mathbf{z}_j^{\mathbf{f}}]^* G_{\xi_i}^{-1}(\omega_j) [z_j^{x_i} - C_i(e^{-i2\pi\omega_j}) \mathbf{z}_j^{\mathbf{f}}] \Bigg] \\ & + \sum_{k=1}^{r} \Bigg[-\frac{1}{2} \sum_{j=0}^{T-1} \log |G_{u_k}(\omega_j)| \\ & - \frac{1}{2} \sum_{j=0}^{T-1} [A_k(e^{-i2\pi\omega_j}) \mathbf{z}_j^{\mathbf{f}}]^* G_{u_k}^{-1}(\omega_j) [A_k(e^{-i2\pi\omega_j}) \mathbf{z}_j^{\mathbf{f}}] \Bigg].\end{aligned}$$

Above, we use $C_i(e^{-i2\pi\omega_j})$ to denote the i^{th} row of matrix $C(e^{-i2\pi\omega_j})$, and $A_k(e^{-i2\pi\omega_j})$ to denote the k^{th} row of matrix $A(e^{-i2\pi\omega_j})$.

Expressing all terms using scalars (rather than vectors), we have

$$\begin{aligned}\ell(X, F; \boldsymbol{\theta}) \simeq \sum_{i=1}^{N} \Bigg[& -\frac{1}{2} \sum_{j=0}^{T-1} \log |G_{\xi_i}(\omega_j)| \\ & - \frac{1}{2} \sum_{j=0}^{T-1} [z_j^{x_i} - (c_{i,01} + c_{i,11} e^{-i2\pi\omega_j}) z_j^{f_1} - ... - (c_{i,0r} + c_{i,1r} e^{-i2\pi\omega_j}) z_j^{f_r}]^* G_{\xi_i}^{-1}(\omega_j) \\ & \quad \times [z_j^{x_i} - (c_{i,01} + c_{i,11} e^{-i2\pi\omega_j}) z_j^{f_1} - ... - (c_{i,0r} + c_{i,1r} e^{-i2\pi\omega_j}) z_j^{f_r}] \Bigg] \\ & + \sum_{k=1}^{r} \Bigg[-\frac{1}{2} \sum_{j=0}^{T-1} \log |G_{u_k}(\omega_j)| \\ & - \frac{1}{2} \sum_{j=0}^{T-1} [z_j^{f_k} - (a_{k1} e^{-i2\pi\omega_j} z_j^{f_1} + ... + a_{kr} e^{-i2\pi\omega_j} z_j^{f_r})]^* G_{u_k}^{-1}(\omega_j) \\ & \quad \times [z_j^{f_k} - (a_{k1} e^{-i2\pi\omega_j} z_j^{f_1} + ... + a_{kr} e^{-i2\pi\omega_j} z_j^{f_r})] \Bigg]\end{aligned}$$

Rearranging terms and plugging in the exact forms of the spectral densities,

$$\begin{aligned}
\ell(X, F; \boldsymbol{\theta}) \simeq \sum_{i=1}^{N} \Bigg[& -\frac{T}{2} \log \Gamma_{\xi_i} \\
& - \frac{1}{2} \Gamma_{\boldsymbol{\xi}_i}^{-1} \sum_{j=0}^{T-1} [z_j^{x_i} - (c_{i,01} + c_{i,11} e^{-i2\pi\omega_j}) z_j^{f_1} - ... - (c_{i,0r} + c_{i,1r} e^{-i2\pi\omega_j}) z_j^{f_r}] \\
& \times [z_j^{x_i} - (c_{i,01} + c_{i,11} e^{-i2\pi\omega_j}) z_j^{f_1} - ... - (c_{i,0r} + c_{i,1r} e^{-i2\pi\omega_j}) z_j^{f_r}]^* \Bigg] \\
& + \sum_{k=1}^{r} \Bigg[-\frac{T}{2} \log \gamma_k \\
& - \frac{1}{2} \gamma_k^{-1} \sum_{j=0}^{T-1} [z_j^{f_k} - (a_{k1} e^{-i2\pi\omega_j} z_j^{f_1} + ... + a_{kr} e^{-i2\pi\omega_j} z_j^{f_r})] \\
& \times [z_j^{f_k} - (a_{k1} e^{-i2\pi\omega_j} z_j^{f_1} + ... + a_{kr} e^{-i2\pi\omega_j} z_j^{f_r})]^* \Bigg],
\end{aligned}$$

which is the final expression we will use for further calculations below.

We now take derivatives with respect to each of the unknown parameters. These yield estimators for the elements of $\boldsymbol{\theta}$ given (i) the latent factors and (ii) the observed series.

2.5.3 Loadings

For a given frequency ω_j, we have that for observations $i = 1, ..., N$, lags $h = 0, 1$, and factors $k = 1, ..., r$,

$$\begin{aligned}
\frac{\partial \ell(X, F; \boldsymbol{\theta})}{\partial c_{i,hk}} = \frac{1}{2} \Gamma_{\xi_i}^{-1} \sum_{j=0}^{T-1} \Bigg[& z_j^{f_k} e^{-i2\pi\omega_j h} [z_j^{x_i} - (c_{i,01} + c_{i,11} e^{-i2\pi\omega_j}) z_j^{f_1} - ... - (c_{i,0r} + c_{i,1r} e^{-i2\pi\omega_j}) z_j^{f_r}]^* \\
& + z_j^{f_k^*} e^{i2\pi\omega_j h} [z_j^{x_i} - (c_{i,01} + c_{i,11} e^{-i2\pi\omega_j}) z_j^{f_1} - ... - (c_{i,0r} + c_{i,1r} e^{-i2\pi\omega_j}) z_j^{f_r}] \Bigg].
\end{aligned}$$

Setting the above to zero defines a system of $2r$ equations for obtaining ML estimators of the dynamic factor loadings for the i^{th} observed variable.

This system of equations (or first-order conditions) is given by

$$\sum_{j=0}^{T-1}\left[z_j^{f_k} z_j^{x_i^*} e^{-i2\pi\omega_j h} + z_j^{f_k^*} z_j^{x_i} e^{i2\pi\omega_j h}\right] =$$
$$\sum_{j=0}^{T-1}\Big[\widehat{c}_{i,01}(z_j^{f_k} z_j^{f_1^*} e^{-i2\pi\omega_j h} + z_j^{f_k^*} z_j^{f_1} e^{i2\pi\omega_j h})$$
$$+ \widehat{c}_{i,11}(z_j^{f_k} z_j^{f_1^*} e^{-i2\pi\omega_j (h-1)} + z_j^{f_k^*} z_j^{f_1} e^{i2\pi\omega_j (h-1)})$$
$$\vdots$$
$$+ \widehat{c}_{i,0r}(z_j^{f_k} z_j^{f_r^*} e^{-i2\pi\omega_j h} + z_j^{f_k^*} z_j^{f_r} e^{i2\pi\omega_j h})$$
$$+ \widehat{c}_{i,1r}(z_j^{f_k} z_j^{f_r^*} e^{-i2\pi\omega_j (h-1)} + z_j^{f_k^*} z_j^{f_r} e^{i2\pi\omega_j (h-1)})\Big].$$

For $i = 1, ..., N$, the left hand side of the above system of equations can be expressed in matrix form as

$$LHS_i^C = \sum_{j=0}^{T-1}\begin{bmatrix} z_j^{f_1} z_j^{x_i^*} + z_j^{f_1^*} z_j^{x_i} \\ z_j^{f_1} z_j^{x_i^*} e^{-i2\pi\omega_j} + z_j^{f_1^*} z_j^{x_i} e^{i2\pi\omega_j} \\ \vdots \\ z_j^{f_r} z_j^{x_i^*} + z_j^{f_r^*} z_j^{x_i} \\ z_j^{f_r} z_j^{x_i^*} e^{-i2\pi\omega_j} + z_j^{f_r^*} z_j^{x_i} e^{i2\pi\omega_j} \end{bmatrix}$$

and the right hand side consists of the matrix

$$RHS^C = \sum_{j=0}^{T-1} \begin{bmatrix} z_j^{f_1} z_j^{f_1^*} + z_j^{f_1^*} z_j^{f_1} & z_j^{f_1} z_j^{f_1^*} e^{i2\pi\omega_j} + z_j^{f_1^*} z_j^{f_1} e^{-i2\pi\omega_j} & \dots \\ z_j^{f_1} z_j^{f_1^*} e^{-i2\pi\omega_j} + z_j^{f_1^*} z_j^{f_1} e^{i2\pi\omega_j} & z_j^{f_1} z_j^{f_1^*} + z_j^{f_1^*} z_j^{f_1} & \dots \\ \vdots & \vdots & \ddots \\ z_j^{f_r} z_j^{f_1^*} + z_j^{f_r^*} z_j^{f_1} & z_j^{f_r} z_j^{f_1^*} e^{i2\pi\omega_j} + z_j^{f_r^*} z_j^{f_1} e^{-i2\pi\omega_j} & \dots \\ z_j^{f_r} z_j^{f_1^*} e^{-i2\pi\omega_j} + z_j^{f_r^*} z_j^{f_1} e^{i2\pi\omega_j} & z_j^{f_r} z_j^{f_1^*} + z_j^{f_r^*} z_j^{f_1} & \dots \end{bmatrix}$$
$$\begin{bmatrix} z_j^{f_1} z_j^{f_r^*} + z_j^{f_1^*} z_j^{f_r} & z_j^{f_1} z_j^{f_r^*} e^{i2\pi\omega_j} + z_j^{f_1^*} z_j^{f_r} e^{-i2\pi\omega_j} \\ z_j^{f_1} z_j^{f_r^*} e^{-i2\pi\omega_j} + z_j^{f_1^*} z_j^{f_r} e^{i2\pi\omega_j} & z_j^{f_1} z_j^{f_r^*} + z_j^{f_1^*} z_j^{f_r} \\ \vdots & \vdots \\ z_j^{f_r} z_j^{f_r^*} + z_j^{f_r^*} z_j^{f_r} & z_j^{f_r} z_j^{f_r^*} e^{i2\pi\omega_j} + z_j^{f_r^*} z_j^{f_r} e^{-i2\pi\omega_j} \\ z_j^{f_r} z_j^{f_r^*} e^{-i2\pi\omega_j} + z_j^{f_r^*} z_j^{f_r} e^{i2\pi\omega_j} & z_j^{f_r} z_j^{f_r^*} + z_j^{f_r^*} z_j^{f_r} \end{bmatrix}$$

post-multiplied by the vector

$$\begin{pmatrix} \widehat{c}_{i,01} \\ \widehat{c}_{i,11} \\ \vdots \\ \widehat{c}_{i,0r} \\ \widehat{c}_{i,1r} \end{pmatrix}.$$

Thus,

$$\begin{pmatrix} \widehat{c}_{i,01} \\ \widehat{c}_{i,11} \\ \vdots \\ \widehat{c}_{i,0r} \\ \widehat{c}_{i,1r} \end{pmatrix} = [RHS^C]^{-1} LHS_i^C$$

are our spectral ML estimators of loadings for the i^{th} observed variable.

More compactly, reverting to vector notation, we have

$$\begin{pmatrix} \widehat{c}_{1,01} & \dots & \widehat{c}_{N,01} \\ \widehat{c}_{1,11} & \dots & \widehat{c}_{N,11} \\ \vdots & \ddots & \vdots \\ \widehat{c}_{1,0r} & \dots & \widehat{c}_{N,0r} \\ \widehat{c}_{1,1r} & \dots & \widehat{c}_{N,1r} \end{pmatrix} = \left\{ \sum_{j=0}^{T-1} \left[\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right)} \right] \right\}^{-}$$

$$\times \sum_{j=0}^{T-1} \left[\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{x}^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{x}^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right)} \right].$$

It will also be useful to re-write the above using a more familiar expression for such an estimator. We define the $2r \times 2$ matrix

$$\dot{\mathbf{Z}}_j^{\mathbf{f}} = \left[\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}, \overline{\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}} \right],$$

and the $N \times 2$ matrix

$$\breve{\mathbf{Z}}_j^{\mathbf{x}} = \left[\mathbf{z}_j^{\mathbf{x}}, \overline{\mathbf{z}_j^{\mathbf{x}}} \right].$$

Then, our estimators have the form

$$\begin{pmatrix} \widehat{c}_{1,01} & \dots & \widehat{c}_{N,01} \\ \widehat{c}_{1,11} & \dots & \widehat{c}_{N,11} \\ \vdots & \ddots & \vdots \\ \widehat{c}_{1,0r} & \dots & \widehat{c}_{N,0r} \\ \widehat{c}_{1,1r} & \dots & \widehat{c}_{N,1r} \end{pmatrix} = \left(\sum_{j=0}^{T-1} \dot{\mathbf{Z}}_j^{\mathbf{f}} \dot{\mathbf{Z}}_j^{\mathbf{f}^*} \right)^{-1} \sum_{j=0}^{T-1} \dot{\mathbf{Z}}_j^{\mathbf{f}} \breve{\mathbf{Z}}_j^{\mathbf{x}^*},$$

which is of course reminiscent of a standard least-squares style formula.

2.5.4 Autoregressive Coefficients

For a given frequency ω_j, we have that for factors $k, l = 1, ..., r$,

$$\frac{\partial \ell(X, F; \boldsymbol{\theta})}{\partial a_{kl}} = \frac{1}{2}\gamma_k^{-1} \sum_{j=0}^{T-1} \Big[z_j^{f_l} e^{-i2\pi\omega_j} [z_j^{f_k} - (a_{k1} e^{-i2\pi\omega_j} z_j^{f_1} + ... + a_{kr} e^{-i2\pi\omega_j} z_j^{f_r})]^* + z_j^{f_l^*} e^{i2\pi\omega_j} [z_j^{f_k} - (a_{k1} e^{-i2\pi\omega_j} z_j^{f_1} + ... + a_{kr} e^{-i2\pi\omega_j} z_j^{f_r})] \Big].$$

Setting the above to zero defines a system of r equations for obtaining the ML estimators of the coefficients in autoregressive equation for the k^{th} factor.

This system of equations (or first-order conditions) is given by

$$\sum_{j=0}^{T-1} \Big[z_j^{f_l} z_j^{f_k^*} e^{-i2\pi\omega_j} + z_j^{f_l^*} z_j^{f_k} e^{i2\pi\omega_j} \Big] = \sum_{j=0}^{T-1} \Big[\widehat{a}_{k1} (z_j^{f_l} z_j^{f_1^*} + z_j^{f_l^*} z_j^{f_1}) + ... + \widehat{a}_{kr} (z_j^{f_l} z_j^{f_r^*} + z_j^{f_l^*} z_j^{f_r}) \Big].$$

For $k = 1, ..., r$, the left hand side of the above system of equations can be expressed in matrix form as

$$LHS_k^A = \sum_{j=0}^{T-1} \begin{bmatrix} z_j^{f_1} z_j^{f_k^*} e^{-i2\pi\omega_j} + z_j^{f_1^*} z_j^{f_k} e^{i2\pi\omega_j} \\ \vdots \\ z_j^{f_r} z_j^{f_k^*} e^{-i2\pi\omega_j} + z_j^{f_r^*} z_j^{f_k} e^{i2\pi\omega_j} \end{bmatrix}$$

and the right hand side consists of the matrix

$$RHS^A = \sum_{j=0}^{T-1} \begin{bmatrix} z_j^{f_1} z_j^{f_1^*} + z_j^{f_1^*} z_j^{f_1} & ... & z_j^{f_1} z_j^{f_r^*} + z_j^{f_1^*} z_j^{f_r} \\ \vdots & \ddots & \vdots \\ z_j^{f_r} z_j^{f_1^*} + z_j^{f_r^*} z_j^{f_1} & ... & z_j^{f_r} z_j^{f_r^*} + z_j^{f_r^*} z_j^{f_r} \end{bmatrix}$$

post-multiplied by the vector

$$\begin{pmatrix} \widehat{a}_{k1} \\ \vdots \\ \widehat{a}_{kr} \end{pmatrix}.$$

Thus,

$$\begin{pmatrix} \widehat{a}_{k1} \\ \vdots \\ \widehat{a}_{kr} \end{pmatrix} = [RHS^A]^{-1} LHS_k^A$$

are our spectral ML estimators of autoregressive coefficients for the k^{th} factor.

More compactly, reverting to vector notation, we have

$$\begin{pmatrix} \widehat{a}_{11} & \dots & \widehat{a}_{r1} \\ \vdots & \ddots & \vdots \\ \widehat{a}_{1r} & \dots & \widehat{a}_{rr} \end{pmatrix} = \left\{ \sum_{j=0}^{T-1} \left[\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} + \overline{\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*}} \right] \right\}^{-1} \sum_{j=0}^{T-1} \left[\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} e^{-i2\pi\omega_j} + \overline{\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} e^{-i2\pi\omega_j}} \right].$$

It will also be useful to re-write the above using a more familiar expression for such an estimator. We define the $r \times 2$ matrix

$$\ddot{\mathbf{Z}}_j^{\mathbf{f}} = \left[\mathbf{z}_j^{\mathbf{f}}, \overline{\mathbf{z}_j^{\mathbf{f}}} \right],$$

and the $r \times 2$ matrix

$$\breve{\mathbf{Z}}_j^{\mathbf{f}} = \left[\mathbf{z}_j^{\mathbf{f}} e^{i2\pi\omega_j}, \overline{\mathbf{z}_j^{\mathbf{f}} e^{i2\pi\omega_j}} \right].$$

Then, our estimators have the form

$$\begin{pmatrix} \widehat{a}_{11} & \dots & \widehat{a}_{r1} \\ \vdots & \ddots & \vdots \\ \widehat{a}_{1r} & \dots & \widehat{a}_{rr} \end{pmatrix} = \left(\sum_{j=0}^{T-1} \ddot{\mathbf{Z}}_j^{\mathbf{f}} \ddot{\mathbf{Z}}_j^{\mathbf{f}^*} \right)^{-1} \sum_{j=0}^{T-1} \ddot{\mathbf{Z}}_j^{\mathbf{f}} \breve{\mathbf{Z}}_j^{\mathbf{f}^*},$$

which is again reminiscent of a standard least-squares style formula.

2.5.5 Variances

For a given frequency ω_j, we have that for observations $i = 1, ..., N$,

$$\frac{\partial \ell(X, F; \boldsymbol{\theta})}{\partial \Gamma_{\xi_i}} = -\frac{T}{2}\Gamma_{\xi_i}^{-1} + \frac{1}{2}\Gamma_{\xi_i}^{-2}\sum_{j=0}^{T-1}\Big[[z_j^{x_i} - (c_{i,01} + c_{i,11}e^{-i2\pi\omega_j})z_j^{f_1} - ... - (c_{i,0r} + c_{i,1r}e^{-i2\pi\omega_j})z_j^{f_r}]$$
$$\times [z_j^{x_i} - (c_{i,01} + c_{i,11}e^{-i2\pi\omega_j})z_j^{f_1} - ... - (c_{i,0r} + c_{i,1r}e^{-i2\pi\omega_j})z_j^{f_r}]^*\Big].$$

We set the above to zero, plug in our ML estimators for the loadings, and obtain

$$\widehat{\Gamma}_{\xi_i} = \frac{1}{T}\sum_{j=0}^{T-1}\Big[[z_j^{x_i} - (\widehat{c}_{i,01} + \widehat{c}_{i,11}e^{-i2\pi\omega_j})z_j^{f_1} - ... - (\widehat{c}_{i,0r} + \widehat{c}_{i,1r}e^{-i2\pi\omega_j})z_j^{f_r}]$$
$$\times [z_j^{x_i} - (\widehat{c}_{i,01} + \widehat{c}_{i,11}e^{-i2\pi\omega_j})z_j^{f_1} - ... - (\widehat{c}_{i,0r} + \widehat{c}_{i,1r}e^{-i2\pi\omega_j})z_j^{f_r}]^*\Big]$$

as our spectral ML estimator for the variance of the i^{th} idiosyncratic component.

In an analagous way, we obtain

$$\widehat{\gamma}_k = \frac{1}{T}\sum_{j=0}^{T-1}\Big[[z_j^{f_k} - (\widehat{a}_{k1}e^{-i2\pi\omega_j}z_j^{f_1} + ... + \widehat{a}_{kr}e^{-i2\pi\omega_j}z_j^{f_r})]$$
$$\times [z_j^{f_k} - (\widehat{a}_{k1}e^{-i2\pi\omega_j}z_j^{f_1} + ... + \widehat{a}_{kr}e^{-i2\pi\omega_j}z_j^{f_r})]^*\Big]$$

as our spectral ML estimator for the variance of the stochastic disturbance term in the k^{th} factor equation.

2.5.6 Some Convergence Results

In this section, we present an unsurprising result; that is, on $\sqrt{T}$ consistency of the maximum likelihood estimator. Indeed, since this is just a textbook property, we state the result only for the loadings estimators. Analogous results hold for maximum likelihood estimators of all the parameters in our model. See, for instance,

Hannan (1973), Dunsmuir and Hannan (1976), and Deistler et al. (1978).[14]

Let $\widehat{\boldsymbol{c}}_i$ denote the $2r \times 1$ vector of estimated loadings for the i^{th} observed variable; that is, $\widehat{\boldsymbol{c}}_i = (\widehat{c}_{i,01}, \widehat{c}_{i,11}, ..., \widehat{c}_{i,0r}, \widehat{c}_{i,1r})'$ for $i = 1, ..., N$.

Proposition 2.4. *Under Assumptions A1-A7, and assuming factors* $\{\mathbf{f}_1, ..., \mathbf{f}_T\}$ *are known, for any* $N \in \mathbb{N}$*, and a given observation* i*, it holds that as* $T \to \infty$*,*

$$\sqrt{T} \left\| \widehat{\boldsymbol{c}}_i - \boldsymbol{c}_i \right\| = O_p(1),$$

uniformly (as defined in Remark 2) with respect to i *for* $i = 1, ..., N$*.*

Proof. See Appendix 2.10.5. □

Remark 3. $\sqrt{T}$-CONSISTENCY. In reality, one should consider not just $\ell(X|F;\boldsymbol{\theta}) + \ell(F;\boldsymbol{\theta})$ but the entire likelihood function $\ell(X;\boldsymbol{\theta})$ for estimation of the parameters. Maximising the former gives rise to a typical least-squares style sampling error which vanishes at rate $\sqrt{T}$ as per Proposition 2.4 whereas maximising the latter results in an *additional* term that is $O_p(N^{-1})$ along with terms that are vanishing with rate faster than $\sqrt{T}$. For results in the time domain, see Barigozzi and Luciani (2022, Lemma 13(i)) and Bai and Li (2016, Theorem 1).

We expect analogous results to hold for our frequency domain estimators. That

[14] Additionally, a textbook treatment is available in Dzhaparidze (1986).

is, for any $i = 1, ..., N$,

$$\begin{aligned}
\widehat{\boldsymbol{c}}_i - \boldsymbol{c}_i = & \left\{ \frac{1}{T} \sum_{j=0}^{T-1} \left[\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right)} \right] \right\}^{-1} \\
& \times \frac{1}{T} \sum_{j=0}^{T-1} \left[\left(\mathbf{z}_j^{\mathbf{f}} z_j^{\zeta_i^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} z_j^{\zeta_i^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right)} \right] \\
& + O_p\left(\frac{1}{N}\right) + o_p\left(\frac{1}{\sqrt{T}}\right) \\
= & \, O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\frac{1}{N}\right) + o_p\left(\frac{1}{\sqrt{T}}\right) \\
= & \, O_p\left(\max\left(\frac{1}{N}, \frac{1}{\sqrt{T}}\right)\right),
\end{aligned}$$

where, say $\sqrt{T}/N \to 0$, we would have that spectral ML estimators are $\sqrt{T}$-consistent.

We do not prove this result or pursue this thread in our study since the extra $O_p\left(N^{-1}\right)$ term that we omit will ultimately be dominated by the $O_p(N^{-1/2})$ error in relation to the WK smoother already established in Section 2.4 above.[15] In other words, the first order conditions based on the likelihood function $\ell(X; \boldsymbol{\theta})$ versus the first order conditions based on $\ell\left(X|F; \boldsymbol{\theta}\right) + \ell\left(F; \boldsymbol{\theta}\right)$ differ only by an amount that will vanish at a rate faster than that of our overall estimation error. $\triangle$

We now make a remark about the formulae for our loadings estimators. The context for this remark will be clear in the next section pertaining to the need for smoothing techniques in non-parametric estimation of spectral density matrices.

Remark 4. LOADINGS ESTIMATORS. As is evident from the expressions in Section 2.5.3, the estimators $\widehat{\boldsymbol{c}}_i$, for $i = 1, ..., N$ are composed primarily of matrices

[15] In fact, as we shall see below, it will also be dominated by an $O_p(N^{-1/2})$ term in the error of our non-parametric estimator of the spectral density matrix. See Proposition 2.5 below. The same spectral density estimator is featured in Forni et al. (2017).

of cross periodograms at each frequency that are summed across frequencies. Clearly, these are "unsmoothed" (see Section 2.6 and, in particular, Section 2.6.2 below) and this is all that is needed to consistently estimate the parameters. As noted in Engle (1974, p.3), the fact that these are unsmoothed periodograms is not a problem; the paper states, "The estimator is consistent, not because each periodogram element approaches its spectral value, but because the sum of the elements approaches the sum of the spectral values which is just the total variance of the variable". Moreover, the variance of the factors is well-behaved due to the ergodicity inherent in the modelling structure in Assumption A2, and consistency follows from the asymptotic independence of the idiosyncratic error and the factors.

2.6 Towards Initialisation: Spectral Density Estimation

With the building blocks for the E step and the M step now in place, we proceed to discuss issues surrounding initialisation of the algorithm. For the exposition in Section 2.6, we consider what happens as $T \to \infty$ with fixed N.

2.6.1 Lag Window Estimator for Spectral Density

We will shortly propose initialisation of our spectral EM procedure using Dynamic PCA, which involves the eigen-decomposition of some non-parametric estimator of the spectral density matrix of $\mathbf{x}_t$. In principle, we could use the periodogram to estimate $G_{\mathbf{x}}(\omega_j)$. Indeed, for a given N, we could compute

$$Per(\omega_j) = \mathbf{z}_j^{\mathbf{x}} \mathbf{z}_j^{\mathbf{x}^*} = \sum_{h=-(T-1)}^{T-1} S_{\mathbf{x}}(h) e^{-i2\pi\omega_j h}$$

where

$$S_{\mathbf{x}}(h) = \frac{1}{T}\sum_{t=|h|+1}^{T} \mathbf{x}_t \mathbf{x}'_{t-|h|} \quad h = 0, ..., T-1, \quad S_{\mathbf{x}}(-h) = S'_{\mathbf{x}}(h),$$

for the Fourier frequencies $\omega_j = j/T$ for $j = 0, ..., T-1$.

The periodogram is asymptotically unbiased. Nevertheless, it exhibits high variability even for very large T, and is in fact an inconsistent estimator of the true spectral density. This may be a surprising result *prima facie* since $Per(\omega_j)$ is essentially the same function (a linear combination) of the sample autocovariances as $G_{\mathbf{x}}(\omega_j)$ is of the theoretical autocovariances. That is, one might be tempted to argue that since $S_{\mathbf{x}}(h)$ is, under general conditions, a consistent estimator of $\Gamma_{\mathbf{x}}(h)$ for any h, a linear combination of $\{S_{\mathbf{x}}(h)\}_{h=-(T-1)}^{T-1}$ would also be a consistent estimator of a linear combination of $\{\Gamma_{\mathbf{x}}(h)\}_{h=-\infty}^{\infty}$; but this is incorrect.

To see why (heuristically), we first note that, under standard conditions on continuity of $\Gamma_{\mathbf{x}}(h)$, the rate at which the squared bias of $S_{\mathbf{x}}(h)$ vanishes will be dominated by the rate at which its variance vanishes in the overall mean-square error computations. So it is the variance of $S_{\mathbf{x}}(h)$, which is $O(1/T)$, that determines the overall rate of convergence. The problem is that while each term in the linear combination is $O(1/T)$, the number of such terms within the overall linear combination is also growing linearly with T. Therein lies the source of the inconsistency. The specific consequence is that the variance of the periodogram does not shrink as T increases. Further details can be found in Priestley (1982).

A common solution to reduce the variance of our estimator as $T \to \infty$ is to use a lag-window (or Blackman-Tukey) estimator which weights the autocovariance sequence so that the number of terms in our linear combination grows at a rate slower than T. Effectively, some of the sample autocovariances are excluded. Of course, this can and does give rise to bias in our estimates. In order to minimise bias, we typically choose to exclude sample autocovariances corresponding to lags that are relatively large (in magnitude). The justification is that continuity of $G_{\mathbf{x}}(\omega_j)$ implies that theoretical autocovariances decay as $|h|$ grows and so excluding large-lag terms is relatively less problematic.

More formally, for some integer $B_T < (T-1)$ such that $B_T \to \infty$ and $B_T/T \to 0$ as $T \to \infty$, we can consider a *truncated* two-sided sequence $S_{\mathbf{x}}(-B_T), ..., S_{\mathbf{x}}(0), ..., S_{\mathbf{x}}(B_T)$ of $(2B_T + 1)$ sample autocovariance matrices. More precisely,

$$S_{\mathbf{x}}(h) = \frac{1}{T} \sum_{t=|h|+1}^{T} \mathbf{x}_t \mathbf{x}'_{t-|h|}, \quad h = 0, ..., B_T, \quad S_{\mathbf{x}}(-h) = S'_{\mathbf{x}}(h),$$

so that B_T is our truncation point.

Such an estimator is referred to as a rectangular (or boxcar) lag-window estimator. The weights are of the form 1 for $|h| \leq B_T$ and 0 for $|h| > B_T$. However, there is no reason why we cannot consider other estimators in which the weighting functions are continuous in h, say, in order to obtain weights that decrease more gradually. As we will observe below, the exact choice of window can indeed affect the asymptotic properties of our estimator, and the rectangular lag window is far from ideal. Indeed, the sharp discontinuities associated with the rectangular lag-window are not a desirable feature for variance reduction. Moreover, we have no guarantee that it generates estimates that are positive definite.

We thus consider a reasonably broad class of windows referred to as "scale parameter" windows. Let us denote our lag-window estimator for $G_{\mathbf{x}}(w_{j'})$, the spectral density matrix of $\mathbf{x}_t$, by $\widetilde{G}^{B_T}_{\mathbf{x}}(w_{j'})$ where B_T is placed in the superscript to emphasise the dependence of the estimator on the size of the window. Specifically, we define

$$\widetilde{G}^{B_T}_{\mathbf{x}}(w_{j'}) = \sum_{h=-(T-1)}^{T-1} S_{\mathbf{x}}(h) \; K\left(\frac{h}{B_T}\right) e^{-i2\pi w_{j'} h}$$

where $w_{j'} = j'/(2B_T+1)$ for $j' = 0, ..., 2B_T$, and $K(\cdot)$ is a fixed weighting function or "kernel" such that

$$K\left(\frac{h}{B_T}\right) = \begin{cases} \mathcal{K}\left(\frac{h}{B_T}\right), & 0 \leq |h| \leq B_T \\ 0, & B_T < |h| \leq T-1, \end{cases}$$

with $\mathcal{K}(\cdot)$ an even, bounded, and continuous function satisfying $\mathcal{K}(0) = 1$.

An increase (or decrease) in the truncation point B_T can be thought of as stretch-

ing (or contracting) the window. Thus, B_T behaves like a scaling parameter. The definitive theory of asymptotic bias, variance and mean-square error for estimators based on scale parameter windows was developed in Parzen (1957). Below, we will generally rely on the more recent treatments presented in Wu and Zaffaroni (2018) and Forni et al. (2017).

Remark 5. Bias-Variance Tradeoff. The effect of truncating and/or reweighting the sample autocovariance sequence is indeed to reduce the variance of our estimator. These manipulations will affect the expected value of the new expression. In general, we can expect the bias to increase. We discuss details in Section 2.6.2 below. △

Remark 6. Revised Grid. We draw attention to the fact that the discretised frequency $w_{j'}$ has necessarily been modified (with respect to ω_j). We discuss details in Section 2.6.3 below. △

Prior to discussing the items alluded to in the foregoing remarks, we borrow (to some extent) from "Assumption 9(i)" in Forni et al. (2017). We present below our assumed regularity conditions for the weighting function.

Assumption A8. *A kernel $K(\cdot)$ is a function with support $[-1, 1]$ such that*

(i) *$K(\cdot)$ is even, bounded, and has the property $K(0) = 1$;*

(ii) *$\int_{-\infty}^{\infty} K^2(u)du < \infty$;*

(iii) *there exists some $\kappa > 0$, the Parzen characteristic exponent of kernel $K(\cdot)$, which represents the largest integer for which*

$$\lim_{u \to 0} \frac{1 - K(u)}{|u|^{\kappa}}$$

is finite and positive;

(iv) *$K(\cdot)$ is Lipschitz continuous; that is, there exists a finite positive constant*

$\overline{M}^K$ *such that*

$$|K(u_1) - K(u_2)| \leq \overline{M}^K |u_1 - u_2|$$

for $u_1, u_2 \in [-1, 1]$.

Remark 7. ASSUMPTION A8(iii). The Parzen characteristic exponent captures the smoothness of the kernel at zero. Indeed, the smoother is the kernel at zero, the larger is the value of κ. For instance, as explained in Andrews (1991), κ equals 1 for the Bartlett kernel (which is triangular in shape with a kink at zero), but 2 for the Parzen, Daniell, General Tukey, Tukey-Hanning, Tukey-Hamming and Bartlett-Priestley kernels. The periodogram, denoted previously by $Per(\omega_j)$, incorporates the rectangular kernel which has a Parzen characteristic exponent of ∞. See Priestley (1982, p. 463). △

Remark 8. ASSUMPTION A8(iv). Part (iv) of our Assumption A8 is stronger than Assumption 9(i)(3) of Forni et al. (2017). In particular, Lipschitz continuity implies bounded variation, and as explained in Liu and Wu (2010)[16], this in turn implies Assumption 9(i)(3) of Forni et al. (2017). While Lipschitz continuity is relatively stronger than the latter, this assumption is not binding in practice. As explained in the context of Assumption 1(b) of Newey and West (1994), several commonly used kernels satisfy this condition. △

2.6.2 Bias-Variance Tradeoff

Lag-window estimators necessitate an inherent trade-off between variance reduction and frequency resolution.[17] The lower is B_T relative to T, the less erratic is our estimator. Indeed, given N, for each frequency $w_{j'} = j'/(2B_T + 1)$ for $j' = 0, ..., 2B_T$, the variance of $\widetilde{G}^{B_T}_{\mathbf{x}}(w_{j'})$ will be $O(B_T/T)$. The cost, however, is

[16] See paragraph immediately following "Condition 3".

[17] The trade-off is somewhat similar, for instance, to that arising when estimating a probability density function via a histogram.

a loss in the ability of our estimator to reveal variation in the spectrum between closely spaced frequencies. In fact, the periodogram may be interpreted as an estimator capable of identifying peaks in the spectrum at the finest possible frequencies, the Fourier frequencies. Lag window methods, which weight the sample autocovariances so as to reduce the contribution from the tail, have the same effect as *smoothing* the periodogram over adjacent frequencies. While this may mitigate against spurious spikes and troughs, there is an inevitable decrease in resolution. As a result, details in the spectrum that are separated by less than $1/(2B_T + 1) > 1/T$ cycles per sampling interval cannot be resolved.

In fact, variance and resolution are affected not just by the size but also the type of lag window. For example, multiplication of the estimated autocovariance sequence in the time domain by a rectangular lag window (say of size B_T) corresponds to convolution in the frequency domain with the Dirichlet spectral smoothing window. The latter features sidelobes that are relatively tall, which is a reflection of the jump discontinuities in the rectangular lag window. This causes spectral *leakage*, which manifests itself through spurious spikes and troughs. In contrast, the triangular lag window, which attenuates end-points of the truncated signal more smoothly, is associated with the Fejér spectral window which features much shorter sidelobes. This reduces spectral leakage. The downside, however, is that the Fejér spectral smoothing window features a relatively broad mainlobe. This causes spectral *smearing*, or loss in resolution. Again, we have a bias-variance tradeoff since one has to strike a balance between mainlobe width and sidelobe height when choosing a suitable window function. There are several possible choices. Nevertheless, we do not consider alternatives. Following the literature, we rely on the triangular lag window. See Forni et al. (2000) and Forni et al. (2017).

To summarise, we need to lower B_T/T to lower the variance of our estimator. In contrast, we need to raise B_T to lower the bias of our estimator. It follows then that choosing B_T such that $B_T/T \to 0$ and $B_T \to \infty$ as $T \to \infty$ (as stated previously) is the key to achieving mean-square convergence.

2.6.3 Coarse Time and Frequency Grids

As mentioned above, in order to consistently estimate the spectral density, we rely on smoothing techniques, which inevitably entail a loss in resolution. In particular, the lag window estimator proposed above is computed not at all T Fourier frequencies but at only $2B_T + 1 < T$ frequencies. Consequently, reverting to the time domain (i.e. via the Inverse DFT), we will be able to estimate the factors $\mathbf{f}_t$ not on the original grid of T time points but on a relatively coarse grid of $2B_T + 1$ points. That is, we will compute some $\hat{\mathbf{f}}_{t'}^W$, where t' denotes a revised temporal index such that $t' \in \mathcal{T}'$ where $\mathcal{T}'$ is a set with cardinality $2B_T + 1$.

Typically, the revised temporal grid is constructed so as to divide the period under consideration into $2B_T + 1$ equally-spaced intervals. That is, while the space between two adjacent points of the original grid is $\Delta t = 1$, for the revised grid, it is $\Delta t' = T/(2B_T + 1)$. Of particular note is that the revised (coarse) grid is in general not synchronous with the original (fine) grid. This is a natural artefact of the interpolative effect of smoothing the periodogram.

Our particular aim, however, will be to discuss pointwise convergence of our estimators. Thus, we need to ensure that there is an overlap between the grids (i.e. for a subset of $2B_T + 1 < T$ points on the original grid). This facilitates a comparison between our estimator and the true value, pointwise in t' for $t' \in \mathcal{T}'$.

To this end, we use the following methodology to construct our revised grid. We set $t' = t_q$ such that

$$t_q \in \left\{ T - (2B_T + 1 - q) \left\lfloor \frac{T}{2B_T + 1} \right\rfloor : q = 1, ..., 2B_T + 1 \right\}$$

denotes our revised temporal index, where the function $\lfloor \cdot \rfloor$ returns the largest integer that is less than or equal to any (real-valued) argument. In other words, we start with the most recent observation, aligning the $(2B_T + 1)^{th}$ point on our revised grid with the T^{th} point on the original grid. We then work backwards aligning each preceding point on our revised grid, with points that are exactly $\lfloor \frac{T}{2B_T+1} \rfloor$ apart on the original grid.

For the frequency domain, we set $j' = j_q$ such that for $j \in \{0, ..., T-1\}$,

$$j_q \in \left\{ \min \left\{ j : \frac{j}{T} \geq \frac{q-1}{2B_T+1} \right\} : q = 1, ..., 2B_T + 1 \right\}$$

denotes our revised frequency index. In other words, we start with frequency zero, aligning the first point on our revised grid with the first point on the original grid. We then work forwards by aligning each consecutive frequency on our revised grid with the smallest available Fourier frequency that is greater than or equal to the latter.

To summarise our notation, we will use

$$\left\{ \left\| \hat{\mathbf{z}}^{\mathbf{f}^W}_{j_q} - \mathbf{z}^{\mathbf{f}}_{j_q} \right\| : q = 1, ..., 2B_T + 1 \right\}, \text{ and}$$

$$\left\{ \left\| \hat{\mathbf{f}}^{W}_{t_q} - \mathbf{f}_{t_q} \right\| : q = 1, ..., 2B_T + 1 \right\}$$

to denote the sets of our factor estimation errors in the frequency and time domains respectively. Typically, our convergence results will be stated either uniformly in j_q (following our weaker definition of uniformity in Remark 2), or pointwise in t_q for $q = 1, ..., 2B_T + 1$. This will be sufficient for our purposes.

Finally, we note that the density of the revised grid, corresponding to our effective sample size, is increasing (with B_T) but its relative density with respect to the original grid, or the original sample size, is decreasing to zero (with B_T/T).

2.6.4 Consistency of Lag Window Estimator

We do not explicitly prove consistency of the lag window estimator proposed for the initialisation of our algorithm because such a proof would not contribute much towards the specific aims of this study. Instead, parallel to the approach in Forni et al. (2009, Assumption 8), we make an additional assumption on the estimate of the cross spectral density between x_{it} and x_{st} for $i, s = 1, ..., N$.

To this end, we leverage results presented in Wu and Zaffaroni (2018). In partic-

ular, the latter finds a $\sqrt{T/(B_T \log B_T)}$ convergence rate. The $\log B_T$ term arises because Wu and Zaffaroni (2018) considers convergence of the extremum over frequencies of the estimation error.[18] Strengthening the usual pointwise consistency result in order to cover convergence of the extremum over frequencies slows the rate from the typical $\sqrt{T/B_T}$ term, for instance in Hallin and Liška (2007), to $\sqrt{T/(B_T \log B_T)}$ in Wu and Zaffaroni (2018). The result is seminal because it applies to a general class of multivariate stationary processes represented by arbitrary measurable functions of *iid* innovations. In particular, no Gaussianity, linearity or strong-mixing assumptions are relied upon. Only primitive assumptions are made; one on finiteness of the p^{th} order moments (for $p > 4$) of the observables, and another on the extent of time dependence of the observables, which involves placing upper bounds on a defined measure referred to as "physical dependence".

Remark 9. EXAMPLES. The concept of physical dependence, first introduced in Wu (2005), has been used in a variety of time series settings in recent years. See, for example, McMurry and Politis (2010), Jirak (2015), Forni et al. (2017), Dette and Gösmann (2018), Wu and Zaffaroni (2018), and Barigozzi et al. (2022). Also see the relatively recent contribution of Zhang and Wu (2021), which extends the dependence measure introduced in Wu (2005) for the high-dimensional setting and considers so-called "nonasymptotic bounds" (Zhang and Wu, 2021, Theorem 4.1 and Corollary 4.4). For succinctness, we prefer to relegate our full exposition of the definition of physical dependence and the restrictions to be imposed on the dependence structure to Appendix 2.10.11. At present, we note simply that there exist many data-generating processes that can be accommodated within the framework of Wu (2005). Indeed, a key reason behind the popularity of the latter is the generality that it admits. Some examples of linear and non-linear processes are mentioned in Liu and Wu (2010, p.1220), and theorems establishing the validity of such examples may be found in Shao and Wu (2007, Section 5). For the purpose of this remark, however, we refer the reader to the excellent overview in

[18] This is what is typically referred to as *uniform* convergence in the literature but we refrain from using this word in this context since we have defined uniformity slightly differently for the purpose of our study in Remark 2.

Wu (2011), which provides not only a detailed list of examples but also a flavour of the theoretical arguments to be made in each case. In general, as Wu (2011) explains, the theory applies to stationary processes of the form $z_t = F(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-1}, ...)$, where $F : [\mathbb{R} \times \mathbb{R} \times ...] \to \mathbb{R}$ is a measurable function, and $\boldsymbol{\epsilon}_t$ is an *iid* stochastic vector process. A sequence of this form is called a physical system (Dette et al., 2020, p.1248), and examples covered thereby include: (i) linear processes, (ii) ARMA processes, (iii) Volterra series, (iv) amplitude-dependent exponential autoregressive processes, (v) non-linear AR processes based on the Clayton copula, (vi) bilinear time series, (vii) threshold AR processes, (viii) ARCH processes, (ix) GARCH processes, (x) random coefficient autoregressive processes, and (xi) non-linear heteroskedastic AR processes. (Wu, 2011, Sections 3–4) △

Below, our assumption borrows from Forni et al. (2017, Proposition 6), which itself builds on results in Wu and Zaffaroni (2018). We believe the assumption is reasonable since its result holds under conditions that are compatible with our modelling framework. Readers who may be interested in the detailed derivations behind the statement of the assumption are referred to Appendix 2.10.11.

For $i, s = 1, ..., N$, we define $\widetilde{g}^{B_T}_{\mathbf{x},is}(w_{j_q})$ and $g_{\mathbf{x},is}(w_{j_q})$ as the $(i,s)^{th}$ elements of $\widetilde{G}^{B_T}_{\mathbf{x}}(w_{j_q})$ and $G_{\mathbf{x}}(w_{j_q})$ respectively.

Assumption A9. *There exist finite positive constants M_{var} and M_{bias} such that*

$$\max_{1\leq i,s\leq N} \mathsf{E}\left[\max_{1\leq q\leq 2B_T+1} \left|\widetilde{g}^{B_T}_{\mathbf{x},is}(w_{j_q}) - g_{\mathbf{x},is}(w_{j_q})\right|^2\right] \leq M_{var}\left(\frac{B_T \log B_T}{T}\right) + M_{bias}\left(\frac{1}{B_T^{2\kappa}}\right),$$

for any $N, T \in \mathbb{N}$.

The first and second terms on the right hand side of Assumption A9 inform us of the rates at which respectively the variance and squared bias of our lag window estimator vanish as T grows without bound (for any N). To fix ideas, say we set the positive integer B_T such that $B_T = \lfloor T^\alpha \rfloor$ for some constant α, where $0 < \alpha < 1$. Then, clearly $B_T \to \infty$ and $B_T/T \to 0$ as $T \to \infty$.

Our Assumption A9 is slightly more general than "Proposition 6" in Forni et al. (2017) on two counts. First, we present our statement in terms of a general

choice of kernel $K(\cdot)$, while Forni et al. (2017) specifically employs the Bartlett lag window. See equation "3.3" therein. This means that in our statements above, the Parzen characteristic exponent κ is not suppressed; whereas in Forni et al. (2017) the value of κ is implicitly fixed to 1. Second, we explicitly display the squared bias term in our "Assumption 9", whereas this term does not appear in "Proposition 6" of Forni et al. (2017). The reason is that "Assumption 10" in Forni et al. (2017) effectively limits the choice of α so as to ensure that the variance term will dominate the squared bias term in determining the consistency result. It is only for this reason that the squared bias term can be suppressed in Forni et al. (2017). We discuss this second point in more detail below.

Given κ, for $0 < \alpha < (2\kappa + 1)^{-1}$, the squared bias term will lead the convergence result in Assumption A9. To see this, let's assume (for ease of exposition) that T^α is an integer so that $\lfloor T^\alpha \rfloor = T^\alpha$. Then, the variance term is

$$\frac{B_T \log B_T}{T} = \frac{T^\alpha \log T^\alpha}{T} = \alpha T^{\alpha-1} \log T,$$

the squared bias term is

$$\frac{1}{B_T^{2\kappa}} = T^{-2\kappa\alpha},$$

and the ratio of these terms evaluates to

$$\alpha T^{(2\kappa+1)\alpha-1} \log T.$$

With a given value of κ, for $0 < \alpha < (2\kappa + 1)^{-1}$, the ratio above will tend to 0 as $T \to \infty$. In other words, the squared bias term will lead the convergence result. On the other hand, for $(2\kappa + 1)^{-1} \leq \alpha < 1$, the ratio above will tend to ∞ as $T \to \infty$, which means that the variance term will lead instead.

We can now see that our exposition is entirely compatible with Forni et al. (2017) wherein the lower bound for α is simply assumed to be larger than or equal to 1/3. This makes sense since (i) the authors employ the Bartlett lag window, for which the value for κ is equal to 1, and (ii) the authors presumably wished to focus simply on rate of variance reduction in their "Proposition 6". As noted previously, our exposition is just slightly more general. For instance, with our

approach, it is easy to observe that if one were to employ, say, the Parzen or Daniell lag windows, for which κ is equal to 2, the appropriate threshold for α would be 1/5.

Restricting our attention to the case that $(2\kappa+1)^{-1} \leq \alpha < 1$, one may compare our assumed rate $O(B_T \log B_T/T)$ with the rate $O(1/T)$ which is standard in the time domain literature concerning estimation of the covariance matrix of the observables. For instance, Forni et al. (2009) has "Assumption 8", wherein it is assumed that $\mathsf{E}\left[(\widehat{\gamma}_{\mathbf{x},is} - \gamma_{\mathbf{x},is})^2\right] < T^{-1}M^{\gamma}$; the terms $\gamma_{\mathbf{x},is}$ and $\widehat{\gamma}_{\mathbf{x},is}$ denote the covariance and estimated covariance between x_{it} and x_{st}, the inequality holds for all $i, s = 1, ..., N$, and M^{γ} is some finite positive constant. By comparison, the extra term B_T in the numerator of our assumed rate, which slows the speed of convergence, is the additional price we pay for requiring initialisation via consistent non-parametric estimation of the spectral density matrix of the observables for a given frequency (and a given N). Moreover, the extra term $\log B_T$ in the numerator, which further slows the speed of convergence, is the additional price for requiring a consistency result that holds uniformly with respect to the frequencies (and for any N).

2.7 Spectral EM Algorithm: Implementation

We now possess all building blocks needed in order to summarise the key steps in implementing the Spectral EM algorithm. Say assumptions A1-A9 hold, and our objective is to estimate the factors and parameters of the model. Then, the first step of our proposed procedure would be to construct a lag window estimator for the spectral density matrix of $\mathbf{x}_t$, denoted by $\widetilde{G}_{\mathbf{x}}^{B_T}(w_{j_q})$, for $q = 1, ..., 2B_T + 1$, with an appropriate choice of kernel. Subsequent steps are as follows.

2.7.1 Initialisation: Factors

We initialise the algorithm using dynamic PCA following Forni et al. (2000), whereby we aim to carry out the eigen-decomposition of our $\widetilde{G}_{\mathbf{x}}^{B_T}(w_{j_q})$. The eigenvectors thus obtained are hereafter referred to as the *dynamic eigenvectors* of $\mathbf{x}_t$.

Specifically, we compute the N (complex orthogonal unit-modulus) eigenvectors of $\widetilde{G}_{\mathbf{x}}^{B_T}(w_{j_q})$ as our estimators of the dynamic eigenvectors of $\mathbf{x}_t$. We retain and gather the first r of these N eigenvectors in an $N \times r$ matrix denoted by $\widetilde{P}_{\mathbf{x}}(w_{j_q})$. This forms our estimate of the first r (complex orthogonal unit-modulus) eigenvectors of the spectral density matrix of the common component, which are themselves contained in the $N \times r$ matrix denoted by $P_{\boldsymbol{\chi}}(w_{j_q})$. We denote the corresponding eigenvalue matrices as $\widetilde{\Lambda}_{\mathbf{x}}(w_{j_q})$ and $\Lambda_{\boldsymbol{\chi}}(w_{j_q})$.

We define the estimator $\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P}$ as

$$\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} = \widetilde{P}_{\mathbf{x}}^{*}(w_{j_q})\mathbf{z}_{j_q}^{\mathbf{x}}/\sqrt{N}$$

as our Dynamic PCA estimator of the DFT of the unobserved factors noting (as mentioned in Section 2.6.3) that we compute the latter for the frequencies indexed as j_q for $q = 1, ..., 2B_T + 1$.

We refer to $\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P}$ as our *pre-estimator* of the DFT of the unobserved factors at frequency j_q for $q = 1, ..., 2B_T+1$, the theoretical properties of which are presented in Section 2.8.1 below.

2.7.2 Initialisation: Parameters

Next, we carry out a round of maximum likelihood estimation conditional on $\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P}$, for $q = 1, ..., 2B_T + 1$, to obtain our pre-estimator of the parameters of the model.

The loadings pre-estimator $\hat{\mathbf{c}}_i^{(0)}$ is obtained by

$$\begin{pmatrix} \hat{c}_{i,01}^{(0)} \\ \hat{c}_{i,11}^{(0)} \\ \vdots \\ \hat{c}_{i,0r}^{(0)} \\ \hat{c}_{i,1r}^{(0)} \end{pmatrix} = \left\{ \sum_{q=1}^{2B_T+1} \left[\left(\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} \tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^{P*}} \otimes \begin{pmatrix} 1 & e^{i2\pi w_{j_q}} \\ e^{-i2\pi w_{j_q}} & 1 \end{pmatrix} \right) + \overline{\left(\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} \tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^{P*}} \otimes \begin{pmatrix} 1 & e^{i2\pi w_{j_q}} \\ e^{-i2\pi w_{j_q}} & 1 \end{pmatrix} \right)} \right] \right\}^{-1}$$
$$\times \sum_{q=1}^{2B_T+1} \left[\left(\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} z_{j_q}^{x_i^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi w_{j_q}} \end{pmatrix} \right) + \overline{\left(\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} z_{j_q}^{x_i^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi w_{j_q}} \end{pmatrix} \right)} \right]$$

for each observed variable $i = 1, ..., N$.

The pre-estimator for the autoregressive coefficients $\hat{\mathbf{a}}_k^{(0)}$ is obtained by

$$\begin{pmatrix} \hat{a}_{k1}^{(0)} \\ \vdots \\ \hat{a}_{kr}^{(0)} \end{pmatrix} = \left\{ \sum_{q=1}^{2B_T+1} \left[\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} \tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^{P*}} + \overline{\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} \tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^{P*}}} \right] \right\}^{-1} \sum_{q=1}^{2B_T+1} \left[\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} \tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^{P*}} e^{-i2\pi w_{j_q}} + \overline{\tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} \tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^{P*}} e^{-i2\pi w_{j_q}}} \right]$$

for each factor $k = 1, ..., r$.

Given the pre-estimators $\hat{\mathbf{c}}_i^{(0)}$ and $\hat{\mathbf{a}}_k^{(0)}$, we estimate the variances by

$$\hat{\Gamma}_{\xi_i}^{(0)} = \frac{1}{2B_T+1} \sum_{q=1}^{2B_T+1} \left[\left[z_{j_q}^{x_i} - \left(\hat{c}_{i,01}^{(0)} + \hat{c}_{i,11}^{(0)} e^{-i2\pi w_{j_q}} \right) \tilde{z}_{j_q}^{f_1^P} - ... - \left(\hat{c}_{i,0r}^{(0)} + \hat{c}_{i,1r}^{(0)} e^{-i2\pi w_{j_q}} \right) \tilde{z}_{j_q}^{f_r^P} \right] \right.$$
$$\left. \times \left[z_{j_q}^{x_i} - \left(\hat{c}_{i,01}^{(0)} + \hat{c}_{i,11}^{(0)} e^{-i2\pi w_{j_q}} \right) \tilde{z}_{j_q}^{f_1^P} - ... - \left(\hat{c}_{i,0r}^{(0)} + \hat{c}_{i,1r}^{(0)} e^{-i2\pi w_{j_q}} \right) \tilde{z}_{j_q}^{f_r^P} \right]^* \right]$$

for $i = 1, ..., N$; and

$$\hat{\gamma}_k^{(0)} = \frac{1}{2B_T+1} \sum_{q=1}^{2B_T+1} \left[\left[\tilde{z}_{j_q}^{f_k^P} - \left(\hat{a}_{k1}^{(0)} e^{-i2\pi w_{j_q}} \tilde{z}_{j_q}^{f_1^P} + ... + \widehat{a}_{kr}^{(0)} e^{-i2\pi w_{j_q}} \tilde{z}_{j_q}^{f_r^P} \right) \right] \right.$$
$$\left. \times \left[\tilde{z}_{j_q}^{f_k^P} - \left(\hat{a}_{k1}^{(0)} e^{-i2\pi w_{j_q}} \tilde{z}_{j_q}^{f_1^P} + ... + \widehat{a}_{kr}^{(0)} e^{-i2\pi w_{j_q}} \tilde{z}_{j_q}^{f_r^P} \right) \right]^* \right]$$

for $k = 1, ..., r$.

To summarise, we are now able to construct the following key terms; that is, for $q = 1, ..., 2B_T + 1$, we obtain

(i) estimates of the transfer functions of the linear filters in our model

$$\widehat{C}^{(0)}(e^{-i2\pi w_{j_q}}) \text{ and } \widehat{A}^{(0)}(e^{-i2\pi w_{j_q}}),$$

by plugging in the computed pre-estimators into the relevant matrices of interest;

(ii) estimates of the spectral density matrices of the idiosyncratic component and of the innovation process in the factor equation,

$$\widehat{G}^{(0)}_{\boldsymbol{\xi}}(w_{j_q}) \text{ and } \widehat{G}^{(0)}_{\mathbf{u}}(w_{j_q}),$$

noting that at all frequencies these are given simply by $\widehat{\Gamma}^{(0)}_{\boldsymbol{\xi}}$ and $\widehat{\Gamma}^{(0)}_{\mathbf{u}}$ respectively;

(iii) an estimate of the spectral density matrix of the factors

$$\widehat{G}^{(0)}_{\mathbf{f}}(w_{j_q}),$$

given by $\left[\widehat{A}^{(0)^{-1}}(e^{-i2\pi w_{j_q}})\right] \widehat{\Gamma}^{(0)}_{\mathbf{u}} \left[\widehat{A}^{(0)^{-1}}(e^{-i2\pi w_{j_q}})\right]^*$; and

(iv) an estimate of the spectral density of the factor estimation error

$$\widehat{W}^{(0)}(w_{j_q}),$$

given by $\left[\widehat{G}^{(0)^{-1}}_{\mathbf{f}}(w_{j_q}) + \widehat{C}^{(0)^*}(e^{-i2\pi w_{j_q}})\widehat{G}^{(0)^{-1}}_{\boldsymbol{\xi}}(w_{j_q})\widehat{C}^{(0)}(e^{-i2\pi w_{j_q}})\right]^{-1}$.

2.7.3 Spectral EM Algorithm

We summarise the steps of our procedure by means of the pseudocode presented in the table below. As regards notation used therein, let $l\left(X, F; \boldsymbol{\theta}\right)$ represent the complete Whittle log-likelihood computed with respect to the coarse grid of $2B_T + 1$ frequencies. That is,

$$\begin{aligned} l\left(X, F; \boldsymbol{\theta}\right) \simeq & -\frac{1}{2}\sum_{q=1}^{2B_T+1} \log|G_{\boldsymbol{\xi}}(w_{j_q})| - \frac{1}{2}\sum_{q=1}^{2B_T+1} \mathbf{z}_j^{\boldsymbol{\xi}^*} G_{\boldsymbol{\xi}}^{-1}(w_{j_q})\mathbf{z}_j^{\boldsymbol{\xi}} \\ & -\frac{1}{2}\sum_{q=1}^{2B_T+1} \log|G_{\mathbf{u}}(w_{j_q})| - \frac{1}{2}\sum_{q=1}^{2B_T+1} \mathbf{z}_j^{\mathbf{u}^*} G_{\mathbf{u}}^{-1}(w_{j_q})\mathbf{z}_j^{\mathbf{u}}. \end{aligned}$$

Accordingly, at iteration $k \geq 0$, and given estimates $\hat{\boldsymbol{\theta}}^{(k)}$, our "Q-function" is given by

$$Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)}) = \mathsf{E}_{\hat{\boldsymbol{\theta}}^{(k)}}\left[l\left(X, F; \boldsymbol{\theta}\right)|X\right].$$

The procedure is as follows. Given, $\hat{\boldsymbol{\theta}}^{(k)}$, we run the WK smoother to compute an estimator of the DFT of the factors. This is then used to compute the our Q-function, which constitutes the E step. Then, the Q-function is maximised with respect to $\boldsymbol{\theta}$ to find $\hat{\boldsymbol{\theta}}^{(k+1)}$, and this constitutes the M step. Once we have convergence (according to a given criterion), we denote the final estimate of the parameters as $\hat{\boldsymbol{\theta}}^{SpEM}$. We end with a final run of the WK smoother to obtain an estimator of the DFT of the factors and convert this via the inverse DFT to get a time domain estimator of the factors denoted by $\hat{\mathbf{f}}_{t_q}^{SpEM}$.

Algorithm: Spectral EM Algorithm

Input : an $N \times T$ matrix of observations X; a threshold ϵ; pre-estimators of parameters, collectively denoted by vector $\hat{\boldsymbol{\theta}}^{(0)}$

Output: an estimator of the parameters $\hat{\boldsymbol{\theta}}^{SpEM}$; an estimator of the factors $\hat{\mathbf{f}}_{t_q}^{SpEM}$; number of iterations until convergence k^{SpEM}.

set $k = 0$;

while $k \geq 0$ **do**

WK Smoother: compute $\mathsf{E}_{\hat{\boldsymbol{\theta}}^{(k)}}\left[\mathbf{z}_{j_q}^{\mathbf{f}}|X\right]$ denoted by $\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{(k)}}$

$$\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{(k)}} = \widehat{W}^{(k)}(w_{j_q})\widehat{C}^{(k)^*}(e^{-i2\pi w_{j_q}})\widehat{G}_{\boldsymbol{\xi}}^{(k)^{-1}}(w_{j_q})\mathbf{z}_{j_q}^{\mathbf{x}},$$

for $q = 1, ..., 2B_T + 1$;

compute additional sufficient statistics

$$\mathsf{E}_{\hat{\boldsymbol{\theta}}^{(k)}}\left[\mathbf{z}_{j_q}^{\mathbf{f}}\mathbf{z}_{j_q}^{\mathbf{f}^*}|X\right] = \hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{(k)}}\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{(k)^*}} + \widehat{W}^{(k)}(w_{j_q}), \text{ and}$$

$$\mathsf{E}_{\hat{\boldsymbol{\theta}}^{(k)}}\left[\mathbf{z}_{j_q}^{\mathbf{f}}\mathbf{z}_{j_q}^{\mathbf{x}^*}|X\right] = \hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{(k)}}\mathbf{z}_{j_q}^{\mathbf{x}^*},$$

for $q = 1, ..., 2B_T + 1$;

E step: compute $Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)})$;

M step: compute $\hat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)})$;

if $\frac{\left|l(X,F;\hat{\boldsymbol{\theta}}^{(k+1)}) - l(X,F;\hat{\boldsymbol{\theta}}^{(k)})\right|}{\frac{1}{2}\left(\left|l(X,F;\hat{\boldsymbol{\theta}}^{(k+1)})\right| + \left|l(X,F;\hat{\boldsymbol{\theta}}^{(k)})\right|\right)} \geq \epsilon$ **then**

reset $k = k + 1$;

else

define $\hat{\boldsymbol{\theta}}^{SpEM} = \hat{\boldsymbol{\theta}}^{(k+1)}$;

define $k^{SpEM} = k + 1$;

WK Smoother: compute $\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{(k+1)}}$ for $q = 1, ..., 2B_T + 1$;

Inverse DFT: compute

$$\hat{\mathbf{f}}_{t_q}^{SpEM} = \frac{1}{\sqrt{2B_T + 1}} \sum_{q=1}^{2B_T+1} \hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{(k+1)}} e^{i2\pi w_{j_q} t_q},$$

for $q = 1, ..., 2B_T + 1$;

break;

end

end

2.8 Spectral EM Algorithm: Asymptotic Properties

In this section, we present convergence results (as $N, T \to \infty$) for our estimators based on the various building blocks outlined in previous sections.

2.8.1 Initialisation: Factors

We have the following result on the DFT of the factors obtained using Dynamic PCA.

Proposition 2.5. *Under Assumptions A1-A9, it holds that as N and $T \to \infty$,*

$$\max_{1 \leq q \leq 2B_T+1} \left\| \tilde{\mathbf{z}}_{j_q}^{\mathbf{f}^P} - \mathbf{z}_{j_q}^{\mathbf{f}} \right\| = O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T^{2\kappa}}}, \sqrt{\frac{B_T \log B_T}{T}} \right) \right).$$

Proof. See Appendix 2.10.6. □

Proposition 2.5 is equivalent to the key result in Forni et al. (2017, Equation 1.7). The rates are effectively the same.

2.8.2 Initialisation: Parameters

We have the following result on the loadings pre-estimators obtained subsequent to an initial round of maximum likelihood estimation given our pre-estimator of the DFT of the factors (as referred to in Section 2.8.1 above).

Proposition 2.6. *Under Assumptions A1-A9, as N and $T \to \infty$, it holds for any $\kappa > 1/2$ that*

$$\left\| \hat{\boldsymbol{c}}_i^{(0)} - \boldsymbol{c}_i \right\| = O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \sqrt{\frac{B_T \log B_T}{T}} \right) \right),$$

uniformly (as defined in Remark 2) with respect to i where $i = 1, ..., N$.

Proof. See Appendix 2.10.7. □

The intuition for Proposition 2.6 is as follows. Our maximum likelihood estimator (given the pre-estimator of the DFT of the factors) is computed over $2B_T + 1$ points. It follows from Proposition 2.4 that the associated error is $O_p(1/\sqrt{B_T})$. Upon combining this with Proposition 2.5, we see that the latter term dominates $O_p(1/\sqrt{B_T^{2\kappa}})$ term therein whenever $\kappa > 1/2$, so the result in Proposition 6 follows. Indeed, for all commonly used lag windows, κ is typically 1 or 2 (see, for instance, Priestley (1982, p. 463)). For brevity, we do not make further mention of κ hereafter.

We return briefly to the discussion in Remark 3 above. It is clear now from Proposition 2.6 that the extra $O_p(1/N)$ term in the estimation error arising as a result of focussing on the Q-function rather than the likelihood function is of no practical importance for this study. It will inevitably be dominated by the $O_p(1/\sqrt{N})$ terms emanating not only from the Dynamic PCA estimation error in Proposition 2.5 but also the WK smoother to be considered in Proposition 2.7 below.

2.8.3 Final Results

We denote our final loadings estimators, obtained when the Spectral EM algorithm is terminated, as

$$\hat{\boldsymbol{c}}_i^{SpEM} = \hat{\boldsymbol{c}}_i^{\left(k^{SpEM}\right)}$$

for $i = 1, ..., N$.

Proposition 2.7. *Under Assumptions A1-A9, as N and $T \to \infty$, it holds that*

$$\left\| \hat{\boldsymbol{c}}_i^{SpEM} - \boldsymbol{c}_i \right\| = O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \sqrt{\frac{B_T \log B_T}{T}} \right) \right),$$

uniformly (as defined in Remark 2) with respect to i for $i = 1, ..., N$.

Proof. See Appendix 2.10.8. □

Next, we define, for $i = 1, ..., N$,

$$\hat{\boldsymbol{c}}_{i,\cdot}^{SpEM}(L)$$

as the the $r \times 1$ column vector containing the i^{th} row of the matrix lag polynomial $C(L)$ estimated using the Spectral EM procedure. Analogously, we have for $q = 1, ..., 2B_T + 1$,

$$\hat{\boldsymbol{c}}_{i,\cdot}^{SpEM}(e^{-i2\pi w_{jq}})$$

as the associated transfer functions.

Then, we can express as the scalar quantity

$$z_{j_q}^{\chi_i} = \boldsymbol{c}_{i,\cdot}'(e^{-i2\pi w_{jq}})\mathbf{z}_{j_q}^{\mathbf{f}}$$

for $q = 1, ..., 2B_T + 1$. Analogously, we define

$$\hat{z}_{j_q}^{\chi_i^{SpEM}} = \hat{\boldsymbol{c}}_{i,\cdot}^{SpEM'}(e^{-i2\pi w_{jq}})\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{SpEM}}$$

as our corresponding estimator of the DFT of the i^{th} common component.

Proposition 2.8. *Under Assumptions A1-A9, as N and $T \to \infty$, it holds that*

$$\left|\hat{z}_{j_q}^{\chi_i^{SpEM}} - z_{j_q}^{\chi_i}\right| = O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \sqrt{\frac{B_T \log B_T}{T}}\right)\right),$$

uniformly (as defined in Remark 2) with respect to i for $i = 1, ..., N$ and $q = 1, ..., 2B_T + 1$.

Proof. See Appendix 2.10.9. □

Finally, we let the scalar quantities obtained via the Inverse DFTs,

$$\chi_{it_s} = \frac{1}{\sqrt{2B_T+1}} \sum_{q=1}^{2B_T+1} z_{j_q}^{\chi_i} e^{i2\pi w_{j_q}}, \text{ and}$$
$$\widehat{\chi}_{it_s}^{SpEM} = \frac{1}{\sqrt{2B_T+1}} \sum_{q=1}^{2B_T+1} \hat{z}_{j_q}^{\chi_i^{SpEM}} e^{i2\pi w_{j_q}}$$

for $i = 1, ..., N$ and $s = 1, ..., 2B_T + 1$ represent the theoretical and estimated values of the common component respectively.[19]

Proposition 2.9. *Under Assumptions A1-A9, as N and $T \to \infty$, it holds that*

$$\left| \widehat{\chi}_{it_s}^{SpEM} - \widehat{\chi}_{it_s} \right| = O_p \left(\max \left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \sqrt{\frac{B_T \log B_T}{T}} \right) \right),$$

pointwise with respect to i for $i = 1, ..., N$ and $s = 1, ..., 2B_T + 1$.

Proof. See Appendix 2.10.10. □

2.8.4 Discussion of Results

In this section, we comment on our finding above that Propositions 2.6 and 2.7 offer the same convergence rate. In other words, we certainly do not claim to have demonstrated an improvement in speed of convergence for our EM algorithm based procedure over and above a principal components based approach. Nevertheless, we believe that our finding is interesting in its own right for the following reasons.

First, specific rates notwithstanding, our work accords an extensive degree of theoretical support to our main precursor, Fiorentini et al. (2018), which albeit ground-breaking in the frequency domain literature for dynamic approximate

[19] We remind the reader that, as discussed in Section 2.6.3, the indexation t_s for $s = 1, ..., 2B_T + 1$ refers to the relatively coarse temporal grid that arises due to the necessary loss in resolution as a result of smoothing techniques required for non-parametric estimation of the spectral density matrix. The notation has been switched from t_q to t_s simply to avoid conflicting with the indexation over the corresponding coarse frequency grid that appears on the right-hand side of the given expressions.

factor models, is completely silent on the question of asymptotic properties. The paper chooses to focus instead on considerations of computational efficiency. In this regard, our study addresses an important lacuna in the literature. Indeed, Fiorentini et al. (2018) remarks:

> [...] an extension of the Doz et al. (2012) analysis that looks at the properties of our algorithm and the resulting ML estimators in approximate factor models in which the cross-sectional dimension is non-negligible relative to the time series dimension would constitute *a very valuable addition*. In fact, a very large number of series might constitute a computational blessing in this framework, the rationale being that for large N the unobservable factors will be consistently estimated by the Kalman-Wiener-Kolmogorov filter [...]. (Fiorentini et al., 2018, p.269, emphasis added)

Second, our study establishes formally that convergence rates obtained at initialisation are at least preserved through our spectral EM procedure. Indeed, the same finding is celebrated in the time domain analogue of our study, Barigozzi and Luciani (2022). The paper states:

> [...] although potentially subject to mis-specification errors, we show that under assumptions similar to those made for PC analysis [...], the EM estimators of the loadings and the factors have *the same rate of consistency* as those of the PC estimators.[20] (Barigozzi and Luciani, 2022, p.6, emphasis added)

In fact, arguably the most influential work in this area for the entire decade prior to Barigozzi and Luciani (2022) was Doz et al. (2012), which was only able to establish a relatively slow rate of consistency of the factors estimated via the EM algorithm with respect to PCA. Contrast, for example, Doz et al. (2012, Proposition 1) with Barigozzi and Luciani (2022, Proposition 3). Indeed, Barigozzi and Luciani (2022, p.38) states, "[...] our result confirms the conjecture by Doz et al. (2012), based on numerical studies, that, in terms of consistency rates, the EM estimator behaves asymptotically as the PC estimator."

[20] PC is the acronym used in Barigozzi and Luciani (2022) for "principal components".

From this perspective, we believe that by establishing rates in line with the benchmark, we have developed a frequency domain counterpart (for at least some of the results) of Barigozzi and Luciani (2022). We believe this is a non-trivial exercise.

Third, the availability (since 2006[21]) of a formally outlined procedure backed by consistency rates in Doz et al. (2012) (albeit not the sharpest possible) led to a resurgence of interest in the use of likelihood-based methods for factor models for high-dimensional time series. We will very shortly outline some reasons for the popularity of such methods, but let us first consider examples from the literature.

With respect to theory, studies that build upon the contributions of Doz et al. (2012) include (but are not limited to) Bai and Li (2012), Ng et al. (2015), Jungbacker and Koopman (2015), Sundberg and Feldmann (2016), Bai and Li (2016), Bai and Liao (2016), and of course, Barigozzi and Luciani (2022). These papers develop the theory for various likelihood-based procedures for factor model structures in high-dimensional settings.

With respect to applications, studies that leverage and, in some cases, extend the approach of Doz et al. (2012) include (but are not limited to):

(i) Reis and Watson (2010), Luciani (2020), and Barigozzi and Luciani (2021) for constructing indices of economic activity;

(ii) Marcellino and Schumacher (2010), Jungbacker et al. (2011), Bańbura and Modugno (2014), and Marcellino and Sivec (2016) for addressing challenges associated with missing or ragged-edge data[22];

(iii) Giannone et al. (2008), Bańbura et al. (2013), Modugno (2013), and Barnett et al. (2016) for nowcasting;

(iv) Bańbura et al. (2015) for computation of conditional forecasts[23];

[21]The results of Doz et al. (2012) were first made publicly available at least as far back as 2006 prior to official publication in *The Review of Economics and Statistics* in 2012. See, for example, the working paper Doz et al. (2006) freely available from the European Central Bank (ECB) website since September 2006.

[22]Datasets with a ragged-edge refer to unbalanced panels that emerge from publication lags of high- and low-frequency indicators.

[23]These are defined as projections of a set of variables of interest on future paths of some other variables in dynamic systems.

(v) Coroneo et al. (2016), Altavilla et al. (2017), Delle Chiaie et al. (2022), and Barigozzi et al. (2021a) for estimating models with block structures;

(vi) Juvenal and Petrella (2015), and Luciani (2015) for impulse response analysis; and

(vii) Linton et al. (2022) for financial risk management.

Our simple point is that there is clearly a vast amount of interest in using the EM algorithm as an estimation strategy with respect to factor models in high-dimensional settings, and it stands to reason that a serious investigation of theoretical properties is of paramount importance.

We naturally segue to our next question of interest. That is, what indeed is the appeal of the EM algorithm (or likelihood-based procedures in general) to researchers and practitioners alike? The answers offered by Barigozzi and Luciani (2022) are insightful:

> The EM algorithm has two main advantages with respect to PC estimation. First, the EM algorithm allows the user to impose restrictions on the model, thus reflecting any prior knowledge about the data. Indeed, as mentioned above, on the one hand, the state-space formulation and the related Kalman smoother in the E-step allow to impose a variety of different dynamics on the states and to deal with data irregularly spaced in time. On the other hand, in the M-step, we can impose restrictions on the parameters. Although these tasks are in principle possible also by means of PC analysis [...[24]], dealing with missing data and implementing parameter constraints via the EM is simpler and more common in the literature [...]. (Barigozzi and Luciani, 2022, p.6–7)

The paper continues:

> Second, the EM algorithm allows us to model the dynamic evolution of the data explicitly. Indeed, we show that when we estimate a DFM on a dataset of US macroeconomic times series, the EM algorithm

[24] References have been omitted.

> produces estimates of the common component that track the dynamics of the observed series better than the PC estimator. This is especially true for those series displaying periods of high persistence and regime changes, like inflation and interest rates, thus suggesting that the EM might be more robust to local deviations from stationarity. (Barigozzi and Luciani, 2022, p.7)

Further, while we relegate the question of efficiency to future study, there is much evidence to suggest that the EM procedure harbours efficiency gains. Certainly in the time domain, for instance, Barigozzi and Luciani (2022) finds that the loadings estimator of the EM algorithm is asymptotically as efficient as the principal components estimator, while the asymptotic covariance of the estimated factors depends on the degree of cross-sectional correlation in the idiosyncratic component. The paper states:

> In particular, if we strengthen the classical assumption of weak cross-correlation to impose sparsity of the idiosyncratic covariance matrix, the EM estimator of the factors becomes more efficient than the PC estimator. (Barigozzi and Luciani, 2022, p.6)

Finally, there is one specific additional benefit accruing to us in the frequency domain, albeit not in the time domain setting of Barigozzi and Luciani (2022). Specifically, when it comes to forecasting, our spectral EM procedure accords us with a useful advantage over-and-above dynamic PCA. Indeed, it is well-known that dynamic PCA produces an estimator of the common component χ_{it} that is consistent but is based on filters that are two-sided, involving not only present and past values but also future values of $\mathbf{x}_t$ (see, for instance, Forni et al., 2000, Equation 9). This creates no problem in the central part of the sample, but the performance of the estimator of χ_{it} deteriorates as t approaches 1 or T. For this reason, dynamic PCA cannot be used for forecasting. Various studies have attempted to address this (Forni et al., 2005, 2015, 2017; Forni and Lippi, 2011). In similar fashion, we too are able to furnish estimators that do not necessitate the use of two-sided filters and thereby may indeed be used for forecasting.

For all the reasons above, we believe that our findings represent a significant

contribution to the literature. Indeed, it is our hope that availability of benchmark rates for the spectral EM procedure will also galvanise analogous studies in the frequency domain where many sophisticated dynamic specifications (see, for example, Fiorentini et al. (2018)) may be accommodated but research remains scarce.

2.9 Simulations

2.9.1 Model Specification

For $t = 1, ..., T$, we simulate data from a single-lag r-factor model:

$$\begin{aligned} \mathbf{x}_t &= C(L)\mathbf{f}_t + \sqrt{\eta} \odot \epsilon_t, \\ A(L)\mathbf{f}_t &= \mathbf{u}_t \text{ with } \mathbf{u}_t \sim iid\ \mathcal{N}(0, I_r), \\ B(L)\epsilon_t &= \nu_t \text{ with } \nu_t \sim iid\ \mathcal{N}(0, \mathcal{T}), \end{aligned}$$

where[25] in particular,

$$\begin{aligned} \mathbf{x}_t &\text{ is the } N\text{-dimensional vector of observables;} \\ \mathbf{f}_t &\text{ is the } r\text{-dimensional vector of unobserved factors;} \\ C(L) &= \begin{bmatrix} \mathbf{c}_{01} + \mathbf{c}_{11}L & \dots & \mathbf{c}_{0r} + \mathbf{c}_{1r}L \end{bmatrix}; \\ A_{kl}(L) &= \begin{cases} 1 - aL & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases} \text{ for } k, l = 1, ..., r; \\ B_{ij}(L) &= \begin{cases} 1 - bL & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \text{ for } i, j = 1, ..., N; \\ \mathcal{T}_{ij} &= \phi^{|i-j|} \text{ for } i, j = 1, ..., N. \end{aligned}$$

We note that $\mathbf{c}_{0k}$ and $\mathbf{c}_{1k}$ are $N \times 1$ vectors of loadings corresponding to the k^{th} factor and its lag. Specifically, for each $i = 1, ..., N$ and $k = 1, .., r$, we choose

[25] We use $\odot$ to represent the Hadamard (or element-wise) product between two vectors.

loadings $c_{i,0k}$ and $c_{i,1k}$ as draws from the $\mathcal{N}(0,1)$ distribution with the restriction that $\left|\frac{c_{i,0k}}{c_{i,1k}}\right| > 1$ in order to guarantee invertibility of the MA polynomial in the equation for the observables.

We set the scalar autoregressive coefficient $a = 0.7$, and draw initial values for the factor processes, $f_{k,0}$ for all k, from the $\mathcal{N}(0, 1/(1-a^2))$ distribution.

The constant η is an $N \times 1$ vector that allows us to control the proportion of the variance of $\mathbf{x}_t$ accounted for by the idiosyncratic component. Suppose we want this noise-to-signal ratio to be ρ. We can impose the latter by setting $\eta_i = \frac{\rho}{1-\rho}(1-b^2)\Gamma_{\boldsymbol{\chi}_i}$ where $\Gamma_{\boldsymbol{\chi}_i}$ is the variance of the i^{th} common component. Specifically, we set $\rho = 0.5$.

The Toeplitz matrix $\mathcal{T}$ allows us to model cross-sectional correlation in the idiosyncratic component through parameter. Serial correlation in the idiosyncratic component is governed by the autoregressive coefficient. We recall that our spectral EM estimation procedure does not explicitly account for cross-sectional and/or serial correlation in the idiosyncratic component. Thus, the parameters ϕ and b allow us to assess the robustness of our estimators with respect to this form of mis-specification.

The procedure is terminated by looking at the convergence statistic

$$\text{CStat}_m = \frac{\left| l(X, F; \hat{\boldsymbol{\theta}}^{(m)}) - l(X, F; \hat{\boldsymbol{\theta}}^{(m-1)}) \right|}{\frac{1}{2}\left(\left| l(X, F; \hat{\boldsymbol{\theta}}^{(m)}) \right| + \left| l(X, F; \hat{\boldsymbol{\theta}}^{(m-1)}) \right| \right)}$$

where $\theta^{(m)}$ refers to the parameter estimates obtained from the m^{th} iteration. We stop after M iterations if $\text{CStat}_M < 10^{-4}$.

2.9.2 Illustrative Example

In this section, we present an illustration of a single run of our procedure. As regards control parameters, we set $N = 100, T = 200, r = 2, \phi = 0.5, b = 0.3$ and estimate common components in the manner outlined in the previous paragraph. The size of our lag window is set to $2 \times \lfloor \sqrt{T} \rfloor$.

For comparison, we also estimate common components using three other approaches, namely, Static Principal Components Analysis or "SPCA" (see, for example, Stock and Watson, 2002a); the time-domain EM algorithm or "DGR" (see Doz et al., 2011), and Dynamic Principal Components Analysis or "DPCA" (see Forni et al., 2000).

We note that in order to facilitate a convenient comparison across all methods, i.e. with reference to the original temporal grid, we compute the common component in a manner analogous to Forni et al. (2000). That is, we obtain the filter $\widehat{K}_i^{SpEM}(L)$ such that $\widehat{K}_i^{SpEM}(e^{-i2\pi w_{jq}})$ is the product of the transfer function of the loadings and the transfer function of the WK smoother. Indeed,

$$\widehat{K}_i^{SpEM}(e^{-i2\pi w_{jq}}) = \hat{\boldsymbol{c}}_{i,\cdot}^{SpEM'}(e^{-i2\pi w_{jq}})\widehat{W}^{SpEM}(w_{j_q})\widehat{C}^{SpEM^*}(e^{-i2\pi w_{jq}})\widehat{G}_{\boldsymbol{\xi}}^{SpEM^{-1}}(w_{j_q})$$

for $q = 1, ..., 2B_T + 1$. Then, we simply convolve $\widehat{K}_i^{SpEM}(L)$ with the data for the i-th series for $i = 1, ..., N$ and this allows for interpolation on the original temporal grid.

We are then able to measure the performance of each of the four estimators with the following mean-square error criterion:

$$\text{MSE} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}(\widehat{\chi}_{it} - \chi_{it})^2}{\sum_{t=1}^{T}\sum_{i=1}^{N}\chi_{it}^2}$$

where $\widehat{\chi}_{it}$ and χ_{it} represent respectively the estimated and true common component for series i at time t. Of course, the lower is MSE, the better.

The graphs below depict the common component estimates for the first 2 series (chosen simply for illustrative purposes).

Figure 2.1: Common Component for Series 1 (SPCA)

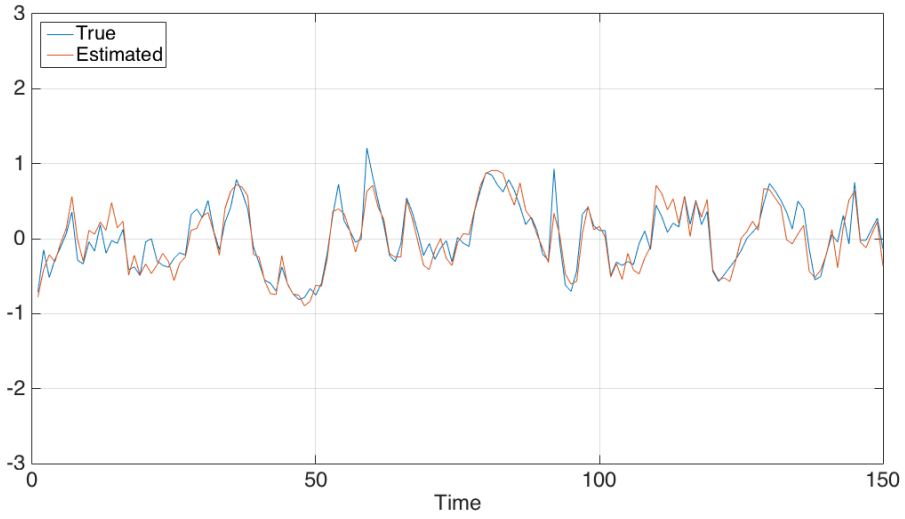


Figure 2.2: Common Component for Series 1 (DGR)

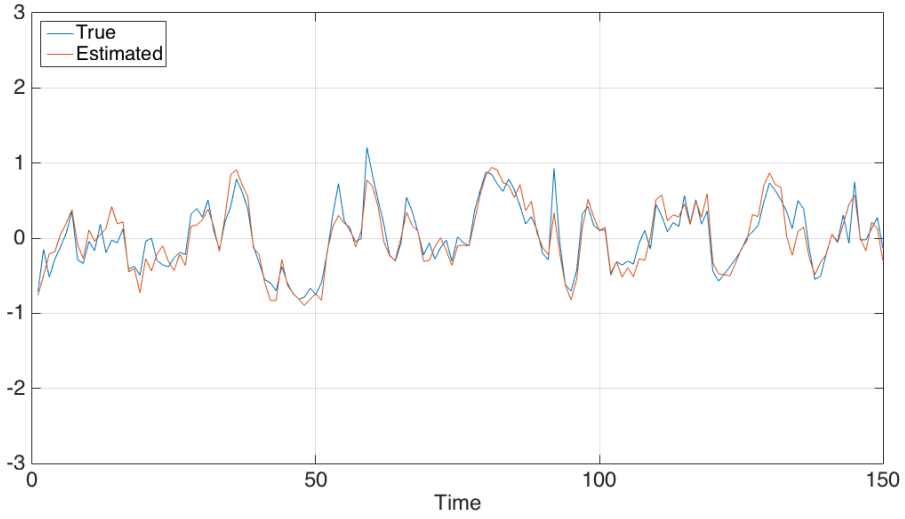


Figure 2.3: Common Component for Series 1 (DPCA)

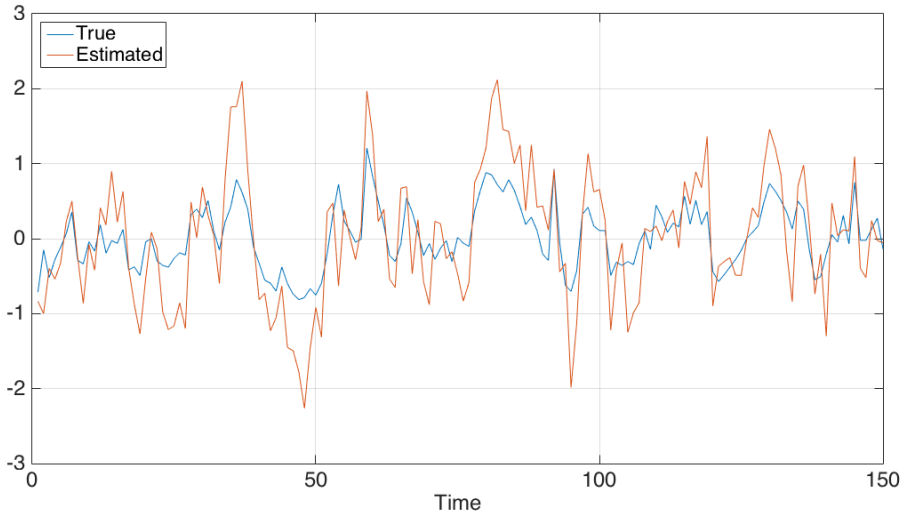


Figure 2.4: Common Component for Series 1 (SpEM)

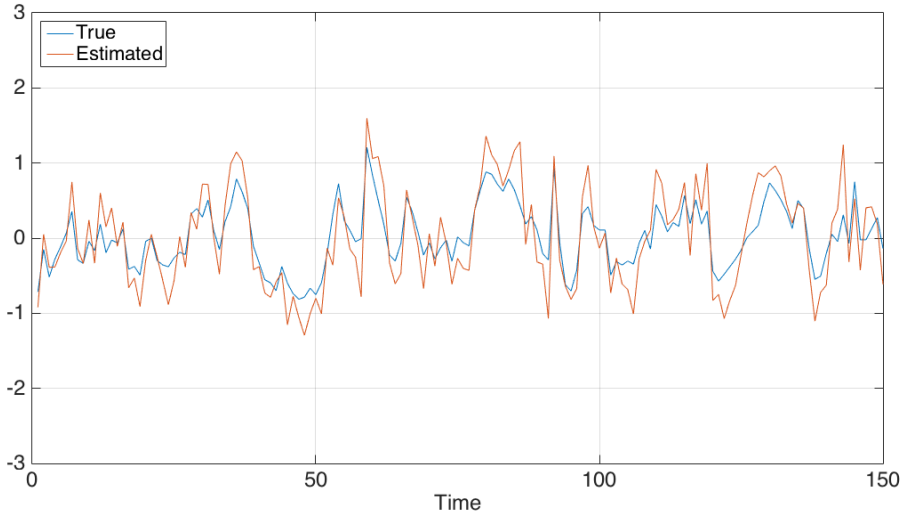


Figure 2.5: Common Component for Series 2 (SPCA)

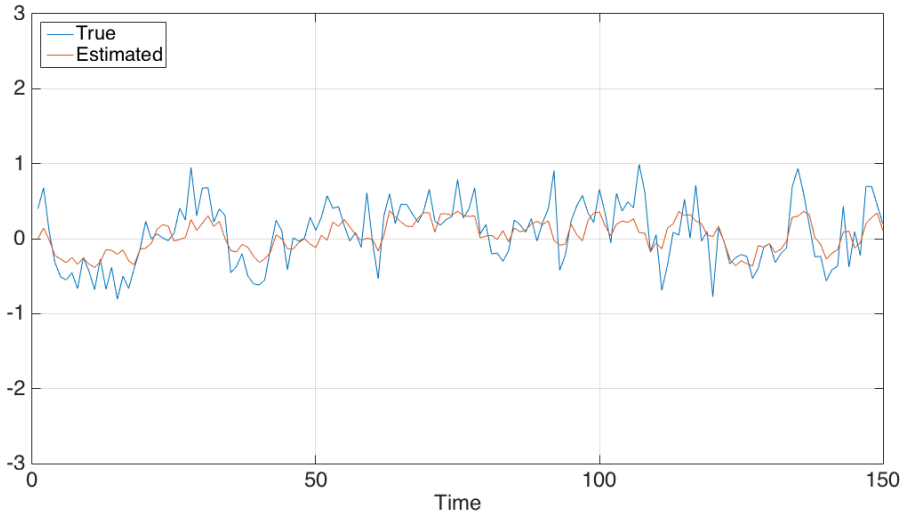


Figure 2.6: Common Component for Series 2 (DGR)

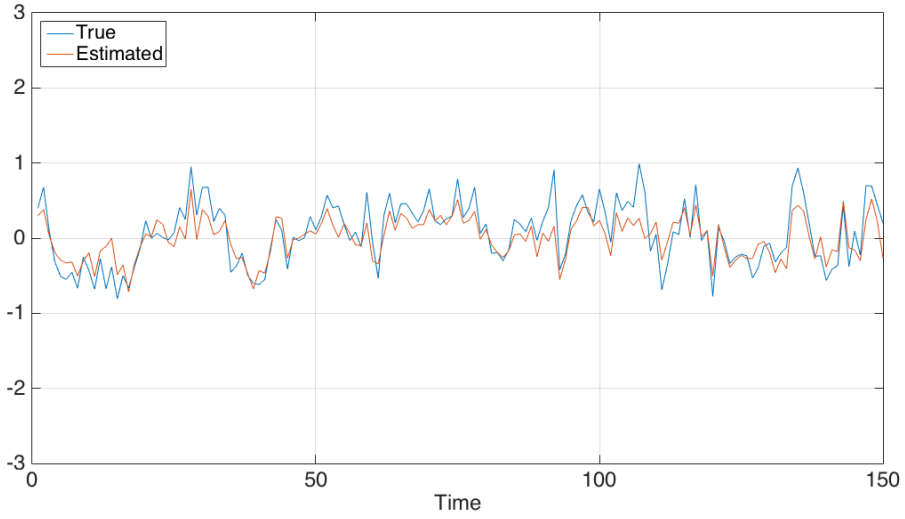


Figure 2.7: Common Component for Series 2 (DPCA)

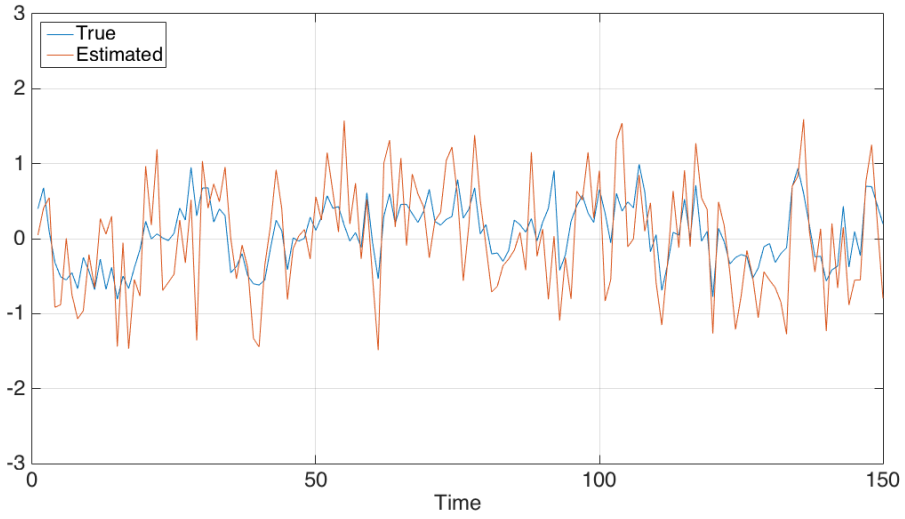
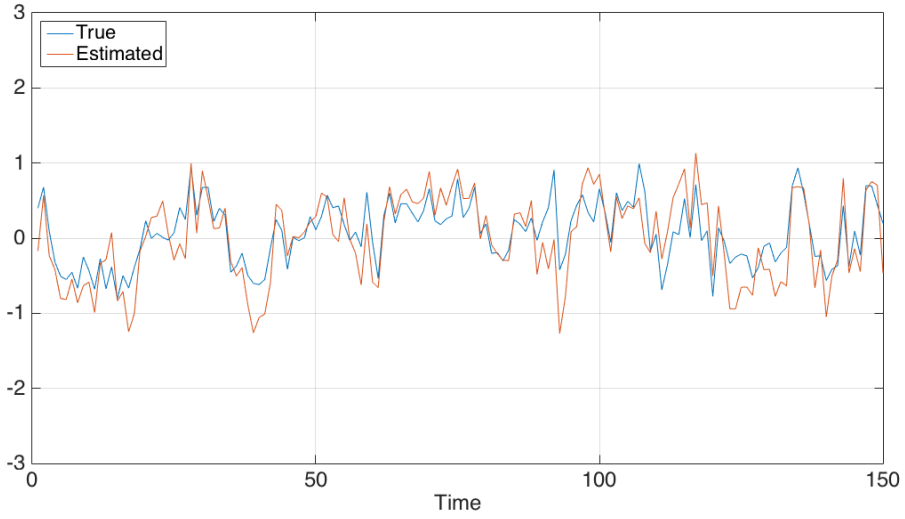


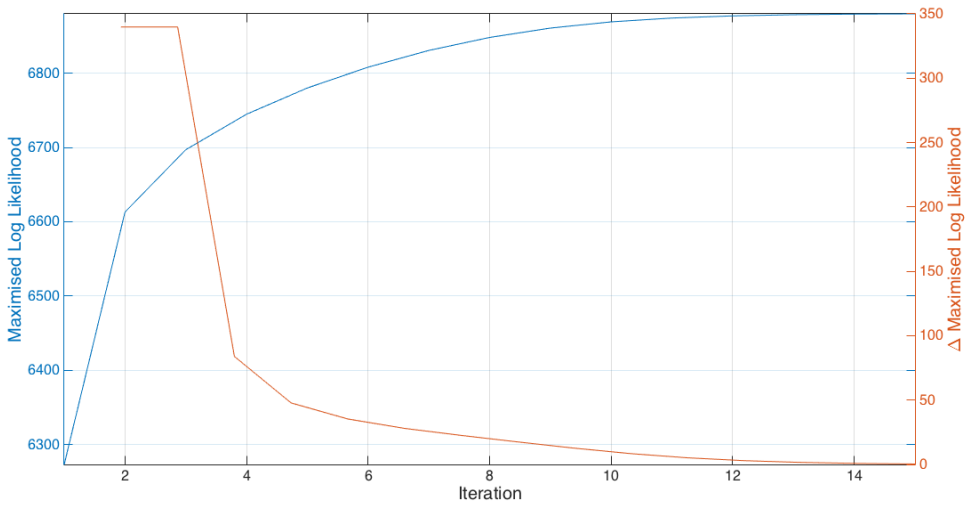
Figure 2.8: Common Component for Series 2 (SpEM)



Results appear to be in accordance with what one might expect. First, it appears that the two time-domain methods outperform the two frequency-domain methods. We believe this is a manifestation of the relative loss in resolution inherent in frequency-domain methods arising due to the need for non-parametric spectral density estimation. That is, for a given T, the effective sample size in the frequency-domain reduces to $2B_T + 1 < T$ points. Second, DGR broadly seems to improve upon SPCA; and correspondingly, SpEM seems to improve upon DPCA. In other words, the two iterative methodologies DGR and SpEM represent gains over and above their one-shot pre-estimators SPCA and DPCA respectively. Of course, our visual analysis is based on only the first two series. By examining the MSE figures we obtained for this illustration, we can also summarise the findings for all N series. Indeed, the MSE for SPCA and DGR obtained were 0.48 and 0.17 respectively and for DPCA and SpEM were 0.65 and 0.48 respectively.

We note as a final point of interest that convergence for SpEM was achieved in 16 iterations in this illustration. The evolution of our convergence statistic is depicted below.

Figure 2.9: Convergence of the SpEM procedure



We conduct a more extensive simulation analysis in the next section.

2.9.3 Monte Carlo Study

Throughout, we let $N \in \{50, 70, 100, 150\}$, $T \in \{70, 100, 150, 200, 300\}$, $r = 2$, $\phi \in \{0, 0.5\}$, and $b \in \{0, 0.3\}$. We consider 100 replications for each individual configuration of these control parameters. Subsequently, we evaluate our estimators by looking at averages of the MSE performance statistic over all repetitions.

Tables 2.1-2.4 report results for cases where innovations are modelled as Gaussian. Specifically, Table 2.1 corresponds to the case in which the estimating model is well-specified; that is, the idiosyncratic component is neither cross-sectionally nor serially correlated ($\phi = 0, b = 0$). On the other hand, Table 2.2 reports results for the case in which the estimating model is mis-specified; that is, the idiosyncratic component is cross-sectionally correlated ($\phi = 0.5$). Table 2.3 considers serial correlation in the idiosyncratic component ($b = 0.3$). Table 2.4 considers both cross-sectional and serial correlation in the idiosyncratic component ($\phi = 0.5, b = 0.3$).

Again, results conform to what we expect. We summarise our findings below.

(i) Given the size of the cross-section N, the quality of SpEM estimates improves as the sample size T increases. SpEM estimates also improve with joint increases in N and T.

(ii) Given N and T, the performance of SpEM deteriorates when we move from considering a well-specified estimating model to a mis-specified one (for example, if we move from Table 2.1 to Table 2.4).

(iii) For any given N and T, the two time-domain procedures SPCA and DGR outperform their frequency-domain counterparts DPCA and SpEM respectively.

(iv) As regards the time-domain procedures, DGR outperforms SPCA for any given N and T. As regards the frequency-domain procedures, SpEM outperforms DPCA for any given N and T. This demonstrates that there are additional gains associated with the iterative procedures over and above their respective one-shot pre-estimators.

Tables 2.5-2.8 report analogous results for the case in which innovations are drawn from the Student's t-distribution with 4 degrees of freedom. We consider this case in order to further assess the performance of SpEM under mis-specification in the context of the true data-generating process following a relatively heavy-tailed distribution. In general, the procedure works as expected. Indeed, the performance of SpEM (along with the other three procedures) deteriorates somewhat relative to the Gaussian case but the overall patterns described above remain broadly unchanged.

Table 2.1: Simulation Results - MSEs
$\phi = 0, b = 0$, Gaussian innovations

SPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.4839	0.4551	0.4368	0.4191	0.4094
N=70	0.4202	0.3798	0.3498	0.3318	0.3236
N=100	0.3283	0.2998	0.2757	0.2562	0.2299
N=150	0.2514	0.2263	0.1900	0.1753	0.1564
DGR					
	T=70	T=100	T=150	T=200	T=300
N=50	0.2006	0.1635	0.1441	0.1309	0.1162
N=70	0.1707	0.1367	0.1147	0.1000	0.0915
N=100	0.1468	0.1153	0.0918	0.0815	0.0693
N=150	0.1290	0.1009	0.0758	0.0667	0.0543
DPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.6826	0.6778	0.6511	0.6415	0.6347
N=70	0.6899	0.6624	0.6438	0.6394	0.6276
N=100	0.6745	0.6586	0.6403	0.6286	0.6193
N=150	0.6763	0.6533	0.6331	0.6236	0.6114
SpEM					
	T=70	T=100	T=150	T=200	T=300
N=50	0.7258	0.6918	0.5503	0.5043	0.5724
N=70	0.5807	0.5007	0.4930	0.4567	0.4519
N=100	0.5295	0.4647	0.4424	0.4403	0.4178
N=150	0.4691	0.4418	0.4239	0.4070	0.3973

Table 2.2: Simulation Results - MSEs
$\phi = 0.5, b = 0$, Gaussian innovations

SPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.5654	0.5195	0.5035	0.4956	0.4854
N=70	0.4811	0.4367	0.4166	0.4091	0.3990
N=100	0.3736	0.3542	0.3253	0.3195	0.3040
N=150	0.3007	0.2653	0.2404	0.2196	0.2075
DGR					
	T=70	T=100	T=150	T=200	T=300
N=50	0.2638	0.2366	0.1992	0.1959	0.1746
N=70	0.2234	0.1812	0.1505	0.1424	0.1293
N=100	0.1764	0.1429	0.1178	0.1045	0.0965
N=150	0.1495	0.1163	0.0942	0.0821	0.0706
DPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.6926	0.6810	0.6564	0.6493	0.6359
N=70	0.6846	0.6680	0.6546	0.6396	0.6312
N=100	0.6796	0.6587	0.6454	0.6340	0.6232
N=150	0.6751	0.6623	0.6385	0.6250	0.6155
SpEM					
	T=70	T=100	T=150	T=200	T=300
N=50	0.6489	0.6466	0.7556	0.5583	0.5247
N=70	0.5962	0.5623	0.5197	0.5286	0.4884
N=100	0.5307	0.4951	0.4830	0.4632	0.4419
N=150	0.4974	0.4672	0.4486	0.4251	0.4063

Table 2.3: Simulation Results - MSEs
$\phi = 0, b = 0.3$, Gaussian innovations

SPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.5303	0.4770	0.4582	0.4271	0.4167
N=70	0.4287	0.3956	0.3590	0.3422	0.3271
N=100	0.3626	0.3276	0.2904	0.2688	0.2465
N=150	0.3128	0.2570	0.2077	0.1885	0.1718
DGR					
	T=70	T=100	T=150	T=200	T=300
N=50	0.2492	0.1967	0.1657	0.1438	0.1278
N=70	0.2112	0.1638	0.1338	0.1145	0.1024
N=100	0.1855	0.1437	0.1117	0.0943	0.0791
N=150	0.1734	0.1286	0.0935	0.0779	0.0633
DPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.6842	0.6663	0.6439	0.6306	0.6254
N=70	0.6735	0.6520	0.6387	0.6283	0.6166
N=100	0.6742	0.6546	0.6279	0.6188	0.6094
N=150	0.6634	0.6481	0.6238	0.6176	0.6050
SpEM					
	T=70	T=100	T=150	T=200	T=300
N=50	0.7057	0.6906	0.5936	0.5656	0.5256
N=70	0.6108	0.5685	0.5170	0.4972	0.4646
N=100	0.5455	0.5269	0.4775	0.4677	0.4390
N=150	0.5095	0.4831	0.4513	0.4352	0.4095

Table 2.4: Simulation Results - MSEs
$\phi = 0.5, b = 0.3$, Gaussian innovations

SPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.5907	0.5493	0.5203	0.5009	0.4962
N=70	0.4829	0.4592	0.4222	0.4247	0.3937
N=100	0.4205	0.3701	0.3476	0.3341	0.3130
N=150	0.3496	0.2926	0.2639	0.2390	0.2197
DGR					
	T=70	T=100	T=150	T=200	T=300
N=50	0.3145	0.2538	0.2242	0.1967	0.1827
N=70	0.2544	0.2004	0.1689	0.1542	0.1372
N=100	0.2159	0.1675	0.1341	0.1190	0.1023
N=150	0.1891	0.1445	0.1124	0.0946	0.0804
DPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.6850	0.6679	0.6534	0.6387	0.6256
N=70	0.6837	0.6570	0.6372	0.6295	0.6226
N=100	0.6684	0.6524	0.6328	0.6246	0.6136
N=150	0.6695	0.6410	0.6272	0.6163	0.6073
SpEM					
	T=70	T=100	T=150	T=200	T=300
N=50	0.9155	0.6861	0.9196	0.6165	0.5575
N=70	0.6309	0.5841	0.5441	0.5528	0.5251
N=100	0.5661	0.5328	0.5097	0.4882	0.4587
N=150	0.5461	0.5004	0.4787	0.4547	0.4380

Table 2.5: Simulation Results - MSEs
$\phi = 0, b = 0$, Student's t innovations

SPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.4077	0.3743	0.3593	0.3360	0.3239
N=70	0.3173	0.2871	0.2811	0.2603	0.2586
N=100	0.2512	0.2275	0.2080	0.1862	0.1754
N=150	0.1991	0.1678	0.1460	0.1269	0.1149
DGR					
	T=70	T=100	T=150	T=200	T=300
N=50	0.1792	0.1501	0.1291	0.1192	0.1083
N=70	0.1463	0.1216	0.1071	0.0956	0.0878
N=100	0.1294	0.1031	0.0880	0.0791	0.0709
N=150	0.1164	0.0958	0.0757	0.0674	0.0587
DPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.7280	0.7260	0.7150	0.7052	0.7089
N=70	0.7283	0.7199	0.7087	0.7120	0.6997
N=100	0.7295	0.7191	0.7061	0.6968	0.6970
N=150	0.7230	0.7118	0.7056	0.6991	0.7004
SpEM					
	T=70	T=100	T=150	T=200	T=300
N=50	0.6831	0.6335	0.6112	0.5954	0.5805
N=70	0.6095	0.5866	0.5554	0.5653	0.5501
N=100	0.5790	0.5667	0.5484	0.5204	0.5192
N=150	0.5646	0.5332	0.5350	0.5135	0.5170

Table 2.6: Simulation Results - MSEs
$\phi = 0.5, b = 0$, Student's t innovations

SPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.4422	0.4247	0.4147	0.4127	0.3902
N=70	0.3753	0.3507	0.3289	0.3186	0.3086
N=100	0.3069	0.2717	0.2586	0.2414	0.2292
N=150	0.2362	0.2019	0.1747	0.1572	0.1468
DGR					
	T=70	T=100	T=150	T=200	T=300
N=50	0.2288	0.1879	0.1787	0.1689	0.1511
N=70	0.1864	0.1533	0.1357	0.1255	0.1147
N=100	0.1558	0.1257	0.1076	0.0999	0.0885
N=150	0.1339	0.1070	0.0896	0.0789	0.0718
DPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.7293	0.7203	0.7241	0.7128	0.7059
N=70	0.7305	0.7187	0.7145	0.7106	0.7037
N=100	0.7302	0.7190	0.7099	0.7039	0.7033
N=150	0.7215	0.7130	0.7064	0.7017	0.7018
SpEM					
	T=70	T=100	T=150	T=200	T=300
N=50	0.6874	0.6410	0.6292	0.6229	0.5991
N=70	0.6515	0.6097	0.5963	0.5932	0.5764
N=100	0.6069	0.5898	0.5675	0.5549	0.5578
N=150	0.5778	0.5585	0.5517	0.5417	0.5409

Table 2.7: Simulation Results - MSEs
$\phi = 0, b = 0.3$, Student's t innovations

SPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.4418	0.4125	0.3849	0.3676	0.3618
N=70	0.3843	0.3404	0.3109	0.2765	0.2796
N=100	0.3309	0.2724	0.2512	0.2187	0.2030
N=150	0.2826	0.2094	0.1675	0.1513	0.1258
DGR					
	T=70	T=100	T=150	T=200	T=300
N=50	0.2163	0.1797	0.1485	0.1319	0.1211
N=70	0.1931	0.1515	0.1232	0.1085	0.0966
N=100	0.1786	0.1308	0.1056	0.0921	0.0779
N=150	0.1625	0.1158	0.0906	0.0772	0.0654
DPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.7323	0.7210	0.7111	0.7051	0.7093
N=70	0.7278	0.7261	0.7110	0.6991	0.7016
N=100	0.7241	0.7123	0.7093	0.7016	0.6997
N=150	0.7238	0.7104	0.7007	0.6918	0.6910
SpEM					
	T=70	T=100	T=150	T=200	T=300
N=50	0.7016	0.6589	0.6302	0.6111	0.6127
N=70	0.6555	0.6476	0.6043	0.5882	0.5803
N=100	0.6255	0.6034	0.5796	0.5672	0.5664
N=150	0.6061	0.5767	0.5555	0.5468	0.5358

Table 2.8: Simulation Results - MSEs
$\phi = 0.5, b = 0.3$, Student's t innovations

SPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.4989	0.4666	0.4558	0.4298	0.4259
N=70	0.4240	0.3985	0.3667	0.3489	0.3431
N=100	0.3686	0.3220	0.2903	0.2649	0.2536
N=150	0.2961	0.2522	0.2070	0.1822	0.1628
DGR					
	T=70	T=100	T=150	T=200	T=300
N=50	0.2766	0.2289	0.2038	0.1790	0.1679
N=70	0.2240	0.1838	0.1599	0.1392	0.1272
N=100	0.1991	0.1556	0.1245	0.1111	0.0995
N=150	0.1724	0.1317	0.1027	0.0885	0.0777
DPCA					
	T=70	T=100	T=150	T=200	T=300
N=50	0.7354	0.7207	0.7121	0.7041	0.7085
N=70	0.7324	0.7184	0.7153	0.7074	0.7008
N=100	0.7323	0.7207	0.7074	0.7038	0.6950
N=150	0.7172	0.7109	0.7059	0.7034	0.6890
SpEM					
	T=70	T=100	T=150	T=200	T=300
N=50	0.8301	0.6879	0.6560	0.6450	0.6378
N=70	0.6894	0.6494	0.6235	0.6202	0.6006
N=100	0.6524	0.6198	0.5984	0.5921	0.5697
N=150	0.6102	0.5898	0.5915	0.5735	0.5480

2.10 Appendix – Proofs for Chapter 2

2.10.1 Proof of Proposition 2.1

First, we note that under Assumptions A1-A3, the matrix $G_{\mathbf{f}}(\omega)$ is positive definite with real, finite and distinct eigenvalues. Next, we observe that

$$\begin{aligned} N^{-1}G_{\boldsymbol{\chi}}(\omega) &= N^{-1}C(e^{-i2\pi\omega})G_{\mathbf{f}}(\omega)C^*(e^{-i2\pi\omega}) \\ &= N^{-1}C(e^{-i2\pi\omega})P_{\mathbf{f}}(\omega)\Lambda_{\mathbf{f}}(\omega)P_{\mathbf{f}}^*(\omega)C^*(e^{-i2\pi\omega}) \\ &= N^{-1}\mathcal{G}_{\boldsymbol{\chi}}(\omega)\mathcal{G}_{\boldsymbol{\chi}}^*(\omega) \end{aligned}$$

where $\mathcal{G}_{\boldsymbol{\chi}}(\omega)$ is defined $\mathcal{G}_{\boldsymbol{\chi}}(\omega) = C(e^{-i2\pi\omega})P_{\mathbf{f}}(\omega)\Lambda_{\mathbf{f}}^{1/2}(\omega)$. Then, for $\omega \in [0,1)$,

$$\begin{aligned} &\mathcal{G}_{\boldsymbol{\chi}}^*(\omega)\mathcal{G}_{\boldsymbol{\chi}}(\omega)/N - \Lambda_{\mathbf{f}}(\omega) \\ &\quad = \Lambda_{\mathbf{f}}^{1/2}(\omega)P_{\mathbf{f}}^*(\omega)C^*(e^{-i2\pi\omega})C(e^{-i2\pi\omega})P_{\mathbf{f}}(\omega)\Lambda_{\mathbf{f}}^{1/2}(\omega)/N - \Lambda_{\mathbf{f}}^{1/2}(\omega)P_{\mathbf{f}}^*(\omega)P_{\mathbf{f}}(\omega)\Lambda_{\mathbf{f}}^{1/2}(\omega) \\ &\quad = \Lambda_{\mathbf{f}}^{1/2}(\omega)P_{\mathbf{f}}^*(\omega)\left[C^*(e^{-i2\pi\omega})C(e^{-i2\pi\omega})/N - I_r\right]P_{\mathbf{f}}(\omega)\Lambda_{\mathbf{f}}^{1/2}(\omega), \end{aligned}$$

so that, by properties of $G_{\mathbf{f}}(\omega)$, there exists a finite positive constant $M^{\boldsymbol{f}}$ such that

$$\begin{aligned} &\sup_{\omega}\|\mathcal{G}_{\boldsymbol{\chi}}^*(\omega)\mathcal{G}_{\boldsymbol{\chi}}(\omega)/N - \Lambda_{\mathbf{f}}(\omega)\| \\ &\qquad \leq \sup_{\omega}\left(\|\Lambda_{\mathbf{f}}^{1/2}(\omega)P_{\mathbf{f}}^*(\omega)\|\right)^2 \sup_{\omega}\|C^*(e^{-i2\pi\omega})C(e^{-i2\pi\omega})/N - I_r\| \\ &\qquad \leq M^{\boldsymbol{f}}\sup_{\omega}\|C^*(e^{-i2\pi\omega})C(e^{-i2\pi\omega})/N - I_r\|. \end{aligned}$$

Then by part (ii) of Assumption A3, clearly

$$\lim_{N\to\infty}\sup_{\omega}\|\mathcal{G}_{\boldsymbol{\chi}}^*(\omega)\mathcal{G}_{\boldsymbol{\chi}}(\omega)/N - \Lambda_{\mathbf{f}}(\omega)\| = 0.$$

We note that $\Lambda_{\mathbf{f}}(\omega)$ is in fact diagonal and has r non-zero, finite and distinct entries due to Assumptions A1, A2, and part (i) of Assumption A3. Since the r largest eigenvalues of $N^{-1}G_{\boldsymbol{\chi}}(\omega) = N^{-1}\mathcal{G}_{\boldsymbol{\chi}}(\omega)\mathcal{G}_{\boldsymbol{\chi}}^*(\omega)$ are of course equal to the r eigenvalues of $N^{-1}\mathcal{G}_{\boldsymbol{\chi}}^*(\omega)\mathcal{G}_{\boldsymbol{\chi}}(\omega)$, we have that as $N \to \infty$, the r largest dynamic

eigenvalues of $\boldsymbol{\chi}_t$ are the same as N times the r finite, non-zero and distinct dynamic eigenvalues of $\mathbf{f}_t$. We thus have our needed conditions on asymptotic divergence (linearly with N) as well as on separation of the dynamic eigenvalues of the common component.

2.10.2 Proof of Proposition 2.2

In this section, we assume the true parameters are known. We prove that the factors can be consistently estimated, as N becomes larger, by the WK smoother.

We begin with the definition of the WK smoother,

$$\mathbf{z}_j^{\mathbf{f}^W} = G_{\mathbf{f}}(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\mathbf{x}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}},$$

for a given frequency ω_j.

Step 1: Application of the Woodbury Formula

We recall from Section 2.4.1.2 that due to the Woodbury formula we can express the WK smoother as

$$\mathbf{z}_j^{\mathbf{f}^W} = W(\omega_j)C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}$$

where $W(\omega_j) = [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}$.

Step 2: Decomposition of $\mathbf{z}_j^{\mathbf{f}^W}$

Plugging back in for $W(\omega_j)$, we have

$$\mathbf{z}_j^{\mathbf{f}^W} = [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}.$$

The next step is to express the inverse, that is the $W(\omega_j)$ term, as the sum of two terms. In order to proceed, we first present the following matrix identities

for square invertible matrices A, B, C and D:

(i) $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$

Proof.

$$\begin{aligned} A^{-1} - B^{-1} &= A^{-1}BB^{-1} - A^{-1}AB^{-1} \\ &= A^{-1}(BB^{-1} - AB^{-1}) \\ &= A^{-1}(B - A)B^{-1} \end{aligned}$$ □

(ii) $(C + D)^{-1} = D^{-1} - (C + D)^{-1}CD^{-1}$

Proof.

$$\begin{aligned} (C + D)^{-1} &= (C + D)^{-1} - D^{-1} + D^{-1} \\ &= (C + D)^{-1}[D - (C + D)]D^{-1} + D^{-1} \\ &\quad \text{(using matrix identity (i))} \\ &= D^{-1} - (C + D)^{-1}CD^{-1} \end{aligned}$$ □

We apply matrix identity (ii) to the previously obtained expression for $\mathbf{z}_j^{\mathbf{f}^W}$, treating $G_{\mathbf{f}}^{-1}(\omega_j)$ as C and $C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})$ as D:

$$\begin{aligned} \mathbf{z}_j^{\mathbf{f}^W} = \Bigg(&[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1} \\ -[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}&G_{\mathbf{f}}^{-1}(\omega_j)[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\Bigg) \\ &\times C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}. \end{aligned}$$

Expanding by bringing the final multiplicative term inside the large parentheses,

we have

$$\mathbf{z}_j^{\mathbf{f}^W} = [C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}$$
$$-[G_{\mathbf{f}}^{-1}(\omega_j)+C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}G_{\mathbf{f}}^{-1}(\omega_j)[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}$$
$$\times C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}.$$

Let us temporarily denote the first term of the previous summation by $z_j^{\mathbf{f}^{W(I_{Temp})}}$ and the second term by $\mathbf{z}_j^{\mathbf{f}^{W(II_{Temp})}}$. That is,

$$\mathbf{z}_j^{\mathbf{f}^W} = \mathbf{z}_j^{\mathbf{f}^{W(I_{Temp})}} - \mathbf{z}_j^{\mathbf{f}^{W(II_{Temp})}},$$

where

$$\mathbf{z}_j^{\mathbf{f}^{W(I_{Temp})}} = [C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}, \text{ and}$$

$$\begin{aligned}\mathbf{z}_j^{\mathbf{f}^{W(II_{Temp})}} &= [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1} \\ &\times G_{\mathbf{f}}^{-1}(\omega_j) \\ &\times [C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1} \\ &\times C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{x}}.\end{aligned}$$

We note that $\mathbf{z}_j^{\mathbf{f}^{W(I_{Temp})}}$ and $\mathbf{z}_j^{\mathbf{f}^{W(II_{Temp})}}$ may each be further split into two separate terms by substituting in for $\mathbf{z}_j^{\mathbf{x}} = C(e^{-i2\pi\omega_j})\mathbf{z}_j^{\mathbf{f}} + \mathbf{z}_j^{\boldsymbol{\zeta}}$ as follows:

First, we see that

$$\mathbf{z}_j^{\mathbf{f}^{W(I_{Temp})}} = \mathbf{z}_j^{\mathbf{f}} + \mathbf{z}_j^{\mathbf{f}^{W(I)}}$$

where

$$\mathbf{z}_j^{\mathbf{f}^{W(I)}} = [C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}.$$

Second, we see that

$$\mathbf{z}_j^{\mathbf{f}^W(II_{Temp})} = \mathbf{z}_j^{\mathbf{f}^W(II)} + \mathbf{z}_j^{\mathbf{f}^W(III)}$$

where

$$\mathbf{z}_j^{\mathbf{f}^W(II)} = [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}G_{\mathbf{f}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{f}}, \text{ and}$$

$$\begin{aligned}\mathbf{z}_j^{\mathbf{f}^W(III)} &= [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1} \\ &\times G_{\mathbf{f}}^{-1}(\omega_j) \\ &\times [C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1} \\ &\times C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}.\end{aligned}$$

Thus, we have an expression for $\mathbf{z}_j^{\mathbf{f}^W}$ involving four terms:

$$\mathbf{z}_j^{\mathbf{f}^W} = \mathbf{z}_j^{\mathbf{f}} + \mathbf{z}_j^{\mathbf{f}^W(I)} - \mathbf{z}_j^{\mathbf{f}^W(II)} - \mathbf{z}_j^{\mathbf{f}^W(III)}.$$

In the subsequent steps below, our goal will be to examine the asymptotic behaviour of each of the terms $\mathbf{z}_j^{\mathbf{f}^W(I)}$, $\mathbf{z}_j^{\mathbf{f}^W(II)}$, and $\mathbf{z}_j^{\mathbf{f}^W(III)}$ as $N \to \infty$.

Step 3: Intermediate Results

In this section, we state three useful lemmas which we will rely upon in the steps that follow.

Lemma 2.10.1. *For $\omega_j = j/T$ where $j = 0, ..., T-1$,*

$$\max_{0\leq j\leq T-1} \mathsf{E}_{\boldsymbol{\theta}} \left[\|N^{-1/2}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}\|^2 \right] = O(1).$$

Proof. We recall that $G_{\boldsymbol{\xi}}^{-1}(\omega_j)$ is an $N \times N$ diagonal matrix and (as above) the i^{th} term of the leading diagonal of this matrix is denoted by $G_{\xi_i}^{-1}(\omega_j)$.

Noting that $C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}$ is an $r \times 1$ vector, we can express its squared spectral norm as

$$\|C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}\|^2 = \sum_{k=1}^{r}\left(\left(\sum_{i=1}^{N}\frac{\overline{C}_{ik}^{j}z_j^{\zeta_i}}{G_{\xi_i}(\omega_j)}\right)\left(\sum_{i=1}^{N}\frac{C_{ik}^{j}\overline{z_j^{\zeta_i}}}{\overline{G_{\xi_i}}(\omega_j)}\right)\right)$$

where a bar on top signifies a complex conjugate (for example, $\overline{C}_{ik}^{j}$ refers to the complex conjugate of the complex scalar C_{ik}^{j}).

Thus, for $\omega_j = j/T$ where $j = 0, ..., T-1$, there exist finite positive constants $M_a^{2.10.1}$, $M_b^{2.10.1}$, and $\overline{M}^{2.10.1}$, independent of ω_j, such that

$$\begin{aligned}
&\mathsf{E}_{\boldsymbol{\theta}}[\|N^{-1/2}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}\|^2] \\
&\quad= N^{-1}\mathsf{E}_{\boldsymbol{\theta}}\left[\sum_{k=1}^{r}\left(\sum_{i=1}^{N}\frac{\overline{C}_{ik}^{j}z_j^{\zeta_i}}{G_{\xi_i}(\omega_j)}\right)\left(\sum_{i=1}^{N}\frac{C_{ik}^{j}\overline{z_j^{\zeta_i}}}{\overline{G_{\xi_i}}(\omega_j)}\right)\right] \\
&\quad\leq N^{-1}\mathsf{E}_{\boldsymbol{\theta}}\left[\sum_{k=1}^{r}\left(\sum_{i=1}^{N}\frac{\overline{C}_{ik}^{j}z_j^{\zeta_i}}{G_{\xi_i}(\omega_j)}\right)\left(\sum_{i=1}^{N}\frac{C_{ik}^{j}\overline{z_j^{\zeta_i}}}{\overline{G_{\xi_i}}(\omega_j)}\right)\right] \\
&\quad= N^{-1}\sum_{k=1}^{r}\mathsf{E}_{\boldsymbol{\theta}}\left[\left(\sum_{i=1}^{N}\sum_{s=1}^{N}\frac{\overline{C}_{ik}^{j}C_{sk}^{j}z_j^{\zeta_i}\overline{z_j^{\zeta_s}}}{G_{\xi_i}(\omega_j)\overline{G_{\xi_s}}(\omega_j)}\right)\right] \\
&\quad\leq N^{-1}rM_a^{2.10.1}\sum_{i=1}^{N}\sum_{s=1}^{N}\mathsf{E}_{\boldsymbol{\theta}}\left[z_j^{\zeta_i}\overline{z_j^{\zeta_s}}\right] \\
&\qquad\text{(by Assumptions A2 and A7)} \\
&\quad= M_b^{2.10.1}N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{N}g_{\boldsymbol{\zeta},is}(\omega_j) \\
&\quad\leq M_b^{2.10.1}M^{\boldsymbol{\zeta}} \\
&\quad\leq \overline{M}^{2.10.1} < \infty \quad \text{(by Assumption A6).}
\end{aligned}$$

Since $\overline{M}^{2.10.1}$ is independent of ω_j, the final result follows. □

For the next two lemmas, we note that the spectral norm of a square Hermitian matrix is equal to the modulus of its largest eigenvalue. For a non-negative definite matrix this is simply equal to the largest eigenvalue.

Lemma 2.10.2. *For $\omega_j = j/T$ where $j = 0, ..., T-1$,*

$$\max_{0 \leq j \leq T-1} \|N[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 = O(1).$$

Proof. We note first that $C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})$ is an $r \times r$ Hermitian non-negative definite matrix. Next, we note that there exist finite positive constants $M_a^{2.10.2}$ and $\overline{M}^{2.10.2}$, independent of ω_j, such that

$$\begin{aligned}
&\|N[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 \\
&\qquad = N^2\left(\lambda_1\{[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\}\right)^2 \\
&\qquad \leq N^2\left(\frac{\lambda_1\{G_{\boldsymbol{\xi}}(\omega_j)\}}{\lambda_r\{C^*(e^{-i2\pi\omega_j})C(e^{-i2\pi\omega_j})\}}\right)^2 \\
&\qquad = \left(\frac{\lambda_1\{G_{\boldsymbol{\xi}}(\omega_j)\}}{\lambda_r\{C^*(e^{-i2\pi\omega_j})C(e^{-i2\pi\omega_j})\}/N}\right)^2 \\
&\qquad \leq M^{\boldsymbol{\xi}} M_a^{2.10.2} \\
&\qquad \leq \overline{M}^{2.10.2} < \infty
\end{aligned}$$

where the penultimate weak inequality follows because (i) the eigenvalue in the numerator is bounded from above due to Assumption A7, and (ii) the denominator is finite and positive due to Assumption A3.

Since $\overline{M}^{2.10.2}$ is independent of ω_j, the final result follows. □

Lemma 2.10.3. *For $\omega_j = j/T$ where $j = 0, ..., T-1$,*

$$\max_{0 \leq j \leq T-1} \|N[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 = O(1).$$

Proof. For the $r \times r$ Hermitian non-negative definite matrices $G_{\mathbf{f}}^{-1}(\omega_j)$ and

$$C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j}),$$

we have

$$\begin{aligned}\|G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})\| \\ \geq \|C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})\|\end{aligned}$$

so that

$$\begin{aligned}\|[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\| \\ \leq \|[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|\end{aligned}$$

It follows by Lemma 2.10.2 that there exists a finite positive constant $\overline{M}^{2.10.3}$, independent of ω_j, such that

$$\|N[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 \leq \overline{M}^{2.10.3} < \infty.$$

Since $\overline{M}^{2.10.3}$ is independent of ω_j, the final result follows. □

Step 4: Asymptotic Analysis of $\mathbf{z}_j^{\mathbf{f}^{W(I)}}$ as $N \to \infty$

We recall the definition of $\mathbf{z}_j^{\mathbf{f}^{W(I)}}$ from Step 2:

$$\mathbf{z}_j^{\mathbf{f}^{W(I)}} = [C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}.$$

Then,

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\theta}}[\|N^{1/2}\mathbf{z}_j^{\mathbf{f}^{W(I)}}\|^2] &= \mathsf{E}_{\boldsymbol{\theta}}[\|N[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}N^{-1/2}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}\|^2] \\ &\leq \|N[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2\mathsf{E}_{\boldsymbol{\theta}}[\|N^{-1/2}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}\|^2] \\ &\quad \text{(by submultiplicativity of the norm)} \\ &\leq \overline{M}^{2.10.2}\overline{M}^{2.10.1} \quad \text{(by Lemmas 2.10.2 and 2.10.1 respectively)} \\ &< \infty.\end{aligned}$$

Since $\overline{M}^{2.10.2}$ and $\overline{M}^{2.10.1}$ are independent of ω_j, we have that

$$\max_{0\le j\le T-1} \mathsf{E}_{\boldsymbol{\theta}}\left[\left\|N^{1/2}\mathbf{z}_j^{\mathbf{f}^{W(I)}}\right\|^2\right] = O(1).$$

Step 5: Asymptotic Analysis of $\mathbf{z}_j^{\mathbf{f}^{W(II)}}$ as $N \to \infty$

We recall the definition of $\mathbf{z}_j^{\mathbf{f}^{W(II)}}$ from Step 2:

$$\mathbf{z}_j^{\mathbf{f}^{W(II)}} = [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}G_{\mathbf{f}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{f}}.$$

Then, there exists a finite positive constant M^{Step5}, independent of ω_j, such that

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\theta}}[\|N\mathbf{z}_j^{\mathbf{f}^{W(II)}}\|^2] &= \mathsf{E}_{\boldsymbol{\theta}}[\|N[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}G_{\mathbf{f}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{f}}\|^2] \\
&\le \|N[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2\mathsf{E}_{\boldsymbol{\theta}}[\|G_{\mathbf{f}}^{-1}(\omega_j)\mathbf{z}_j^{\mathbf{f}}\|^2] \\
&\quad \text{(by submultiplicativity of the norm)} \\
&\le \|N[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 M^{Step5} \\
&\quad \text{(by Assumptions A2-A4)} \\
&\le \overline{M}^{2.10.3} M^{Step5} \quad \text{(by Lemma 2.10.3)} \\
&< \infty.
\end{aligned}$$

Since $\overline{M}^{2.10.3}$ and M^{Step5} are independent of ω_j, we have that

$$\max_{0\le j\le T-1} \mathsf{E}_{\boldsymbol{\theta}}\left[\left\|N\mathbf{z}_j^{\mathbf{f}^{W(II)}}\right\|^2\right] = O(1).$$

Step 6: Asymptotic Analysis of $\mathbf{z}_j^{\mathbf{f}^{W(III)}}$ as $N \to \infty$

We recall the definition of $\mathbf{z}_j^{\mathbf{f}^{W(III)}}$ from Step 2 below:

$$\begin{aligned}\mathbf{z}_j^{\mathbf{f}^{W(III)}} &= [G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1} \\ &\times G_{\mathbf{f}}^{-1}(\omega_j) \\ &\times [C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1} \\ &\times C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}.\end{aligned}$$

Then, there exists a finite positive constant M^{Step6}, independent of ω_j, such that

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\theta}}[\|N^{3/2}\mathbf{z}_j^{\mathbf{f}^{W(III)}}\|^2] &\leq \|N[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\xi}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 \\ &\times \|G_{\mathbf{f}}^{-1}(\omega_j)\|^2 \\ &\times \|N[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 \\ &\times \mathsf{E}_{\boldsymbol{\theta}}[\|N^{-1/2}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}\|^2] \\ &\quad \text{(by submultiplicativity of the norm)} \\ &\leq \|N[G_{\mathbf{f}}^{-1}(\omega_j) + C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 \\ &\times M^{Step6} \\ &\times \|N[C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)C(e^{-i2\pi\omega_j})]^{-1}\|^2 \\ &\times \mathsf{E}_{\boldsymbol{\theta}}[\|N^{-1/2}C^*(e^{-i2\pi\omega_j})G_{\boldsymbol{\xi}}^{-1}(\omega_j)\mathbf{z}_j^{\boldsymbol{\zeta}}\|^2] \\ &\quad \text{(by Assumption A2)} \\ &\leq \overline{M}^{2.10.3} \times M^{Step6} \times \overline{M}^{2.10.2} \times \overline{M}^{2.10.1} \\ &\quad \text{(by Lemmas 2.10.3, 2.10.2, and 2.10.1)} \\ &< \infty.\end{aligned}$$

Since $\overline{M}^{2.10.3}$, M^{Step6}, $\overline{M}^{2.10.2}$, and $\overline{M}^{2.10.1}$ are independent of ω_j, we have that

$$\max_{0\leq j\leq T-1} \mathsf{E}_{\boldsymbol{\theta}}\left[\left\|N^{3/2}\mathbf{z}_j^{\mathbf{f}^{W(III)}}\right\|^2\right] = O(1).$$

Step 7: Asymptotic Analysis of $\mathbf{z}_j^{\mathbf{f}^W}$ as $N \to \infty$

We recall from Step 2 that

$$\mathbf{z}_j^{\mathbf{f}^W} - \mathbf{z}_j^{\mathbf{f}} = \mathbf{z}_j^{\mathbf{f}^W(I)} - \mathbf{z}_j^{\mathbf{f}^W(II)} - \mathbf{z}_j^{\mathbf{f}^W(III)}.$$

Given Steps 4, 5, and 6 above, we notice that, as far as convergence as $N \to \infty$ is concerned, the leading term on the right-hand side is $\mathbf{z}_j^{\mathbf{f}^W(I)}$. It follows that

$$\max_{0 \le j \le T-1} \mathsf{E}_{\boldsymbol{\theta}} \left[N \left\| \mathbf{z}_j^{\mathbf{f}^W} - \mathbf{z}_j^{\mathbf{f}} \right\|^2 \right] = O(1).$$

2.10.3 Proof of Theorem 2.1

Given the DFT of a sequence of time series observations, one may exactly recover the original sequence using the inverse DFT. The inverse relationship between the DFT and the inverse DFT for our time series is given by:

$$\mathbf{f}_t^W = \frac{1}{\sqrt{T}} \sum_{j=0}^{T-1} \mathbf{z}_j^{\mathbf{f}^W} e^{i2\pi\omega_j t}$$

for $t = 1, ..., T$. We recall the proof of this well-known result below.[26]

To prove this relationship, we substitute for $\mathbf{z}_j^{\mathbf{f}^W}$ into the right hand side. We have that for any $t, \tau = 1, ..., T$,

$$\begin{aligned}
\frac{1}{\sqrt{T}} \sum_{j=0}^{T-1} \mathbf{z}_j^{\mathbf{f}^W} e^{i2\pi\omega_j t} &= \frac{1}{\sqrt{T}} \sum_{j=0}^{T-1} \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \mathbf{f}_\tau^W e^{-i2\pi\omega_j \tau} \right) e^{i2\pi\omega_j t} \\
&= \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{f}_\tau^W \sum_{j=0}^{T-1} e^{-i2\pi\omega_j (\tau - t)} \\
&= \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{f}_\tau^W \times \begin{cases} T, & \tau = t \\ 0, & \tau \neq t \end{cases} \\
&= \mathbf{f}_t^W
\end{aligned}$$

where the penultimate line follows due to the orthogonality property of complex exponentials.

By the same reasoning, it also holds that

$$\mathbf{f}_t = \frac{1}{\sqrt{T}} \sum_{j=0}^{T-1} \mathbf{z}_j^{\mathbf{f}} e^{i2\pi\omega_j t}$$

for $t = 1, ..., T$.

[26] See for example Bracewell (2000).

Then, there exists a finite positive constant $\overline{M}^{\mathbf{f}^W}$, independent of ω_j, such that

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\theta}}\left[N\left\|\mathbf{f}_t^W-\mathbf{f}_t\right\|^2\right] &= N\mathsf{E}_{\boldsymbol{\theta}}\left[\frac{1}{T}\left\|\sum_{j=0}^{T-1}(\mathbf{z}_j^{\mathbf{f}^W}-\mathbf{z}_j^{\mathbf{f}})e^{i2\pi\omega_j t}\right\|^2\right] \\
&= N\frac{1}{T}\mathsf{E}_{\boldsymbol{\theta}}\left[\sum_{k=1}^{r}\left(\sum_{j=0}^{T-1}(z_j^{f_k^W}-z_j^{f_k})e^{i2\pi\omega_j t}\right)\left(\sum_{j'=0}^{T-1}\overline{(z_j^{f_k^W}-z_j^{f_k})e^{i2\pi\omega_{j'} t}}\right)\right] \\
&= N\frac{1}{T}\sum_{k=1}^{r}\sum_{j=0}^{T-1}\sum_{j'=0}^{T-1}\mathsf{E}_{\boldsymbol{\theta}}\left[(z_j^{f_k^W}-z_j^{f_k})\overline{(z_{j'}^{f_k^W}-z_{j'}^{f_k})}e^{i2\pi(\omega_j-\omega_{j'})t}\right] \\
&= N\frac{1}{T}\sum_{k=1}^{r}\sum_{j=0}^{T-1}\mathsf{E}_{\boldsymbol{\theta}}\left[\left|(z_j^{f_k^W}-z_j^{f_k})\right|^2\right] \\
&\qquad \text{(by orthogonality of the complex exponential at different frequencies)} \\
&= N\frac{1}{T}\sum_{j=0}^{T-1}\mathsf{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{z}_j^{\mathbf{f}^W}-\mathbf{z}_j^{\mathbf{f}}\right\|^2\right] \\
&\leq \max_{0\leq j\leq T-1}\mathsf{E}_{\boldsymbol{\theta}}\left[N\left\|\mathbf{z}_j^{\mathbf{f}^W}-\mathbf{z}_j^{\mathbf{f}}\right\|^2\right] \\
&\leq \overline{M}^{\mathbf{f}^W} < \infty,
\end{aligned}$$

by Proposition 2.2. This completes the proof of Theorem 2.1.

2.10.4 Proof of Proposition 2.3

Part (i)

We recall that by definition,

$$\mathsf{E}_{\boldsymbol{\theta}}[\mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{f}^*}|\mathcal{D}_T] = \mathbf{z}_j^{\mathbf{f}^W}\mathbf{z}_j^{\mathbf{f}^{W^*}} + W(\omega_j).$$

for a given j.

We recall some results. As $N\to\infty$,

(i) $\|W(\omega_j)\| = O(N^{-1})$ (by Lemma 2.10.3);

(ii) $\|\mathbf{z}_j^{\mathbf{f}}\| = O_p(1)$ (by Assumption A2);

for each of the Fourier frequencies ω_j. Thus,

$$\begin{aligned} E_\theta[\mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{f}^*}|\mathcal{D}_T] &= \left(\mathbf{z}_j^{\mathbf{f}} + O_p(N^{-1/2})\right)\left(\mathbf{z}_j^{\mathbf{f}} + O_p(N^{-1/2})\right)^* + W(\omega_j) \\ &\quad \text{(by Proposition 2.2} \\ &= \mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{f}^*} + O_p(N^{-1/2}). \end{aligned}$$

This completes the proof of part (i) of Proposition 2.3.

Part (ii)

For a given j,

$$\begin{aligned} E_\theta[\mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{x}^*}|\mathcal{D}_T] &= E_\theta[\mathbf{z}_j^{\mathbf{f}}|\mathcal{D}_T]\mathbf{z}_j^{\mathbf{x}^*} \\ &= \mathbf{z}_j^{\mathbf{f}^W}\mathbf{z}_j^{\mathbf{x}^*} \\ &= (\mathbf{z}_j^{\mathbf{f}} + O_p(N^{-1/2}))\mathbf{z}_j^{\mathbf{x}^*} \\ &\quad \text{(by Proposition 2.2)} \\ &= \mathbf{z}_j^{\mathbf{f}}\mathbf{z}_j^{\mathbf{x}^*} + O_p(N^{-1/2}) \end{aligned}$$

since $\|\mathbf{z}_j^{\mathbf{x}}\| = O_p(1)$ due to Assumptions A1-A2.

This completes the proof of part (ii) of Proposition 2.3.

2.10.5 Proof of Proposition 2.4

We recall that the i^{th} row, for $i = 1, ..., N$, in the DFT of our model is

$$
\begin{aligned}
z_j^{x_i} &= C_i(e^{-i2\pi\omega_j})\mathbf{z}_j^{\mathbf{f}} + z_j^{\zeta_i} \\
&= \begin{pmatrix} c_{i,01} & c_{i,11} & c_{i,02} & c_{i,12} & \dots & c_{i,0r} & c_{i,1r} \end{pmatrix}
\begin{pmatrix}
1 & 0 & \dots & 0 \\
e^{-i2\pi\omega_j} & 0 & \dots & 0 \\
0 & 1 & \dots & 0 \\
0 & e^{-i2\pi\omega_j} & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & 1 \\
0 & 0 & \dots & e^{-i2\pi\omega_j}
\end{pmatrix}
\begin{pmatrix} z_j^{f_1} \\ z_j^{f_2} \\ \vdots \\ z_j^{f_r} \end{pmatrix} + z_j^{\zeta_i} \\
&= \begin{pmatrix} c_{i,01} & c_{i,11} & c_{i,02} & c_{i,12} & \dots & c_{i,0r} & c_{i,1r} \end{pmatrix}
\left[\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right] + z_j^{\zeta_i}
\end{aligned}
$$

for a given j, where $\omega_j = j/T$ and $j = 0, ..., T-1$.

Then, the complex conjugate transpose is

$$
z_j^{x_i^*} = \left[\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi w_j} \end{pmatrix} \right]^*
\begin{pmatrix} c_{i,01} \\ c_{i,11} \\ c_{i,02} \\ c_{i,12} \\ \vdots \\ c_{i,0r} \\ c_{i,1r} \end{pmatrix} + z_j^{\zeta_i^*}
$$

since the loadings coefficients are all real.

Let us define the $2r \times N$ matrix of all loadings coefficients $\mathcal{C}$ as

$$\mathcal{C} = \begin{pmatrix} c_{1,01} & \dots & c_{N,01} \\ c_{1,11} & \dots & c_{N,11} \\ c_{1,02} & \dots & c_{N,02} \\ c_{1,12} & \dots & c_{N,12} \\ \vdots & \ddots & \vdots \\ c_{1,0r} & \dots & c_{N,0r} \\ c_{1,1r} & \dots & c_{N,1r} \end{pmatrix}.$$

The definition of $\breve{\mathbf{Z}}_j^{\mathbf{x}} = \left[\mathbf{z}_j^{\mathbf{x}}, \overline{\mathbf{z}_j^{\mathbf{x}}}\right]$ means that

$$\begin{aligned} \breve{\mathbf{Z}}_j^{\mathbf{x}^*} &= \begin{bmatrix} \mathbf{z}_j^{\mathbf{x}^*} \\ \mathbf{z}_j^{\mathbf{x}'} \end{bmatrix} \\ &= \begin{bmatrix} \left[\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}\right]^* \mathcal{C} + \mathbf{z}_j^{\boldsymbol{\zeta}^*} \\ \left[\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}\right]' \mathcal{C} + \mathbf{z}_j^{\boldsymbol{\zeta}'} \end{bmatrix} \end{aligned}$$

so that the product $\dot{\mathbf{Z}}_j^{\mathbf{f}} \breve{\mathbf{Z}}_j^{\mathbf{x}^*}$ is

$$\begin{aligned} \dot{\mathbf{Z}}_j^{\mathbf{f}} \breve{\mathbf{Z}}_j^{\mathbf{x}^*} &= \left[\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}, \overline{\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}}\right] \begin{bmatrix} \left[\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}\right]^* \mathcal{C} + \mathbf{z}_j^{\boldsymbol{\zeta}^*} \\ \left[\mathbf{z}_j^{\mathbf{f}} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}\right]' \mathcal{C} + \mathbf{z}_j^{\boldsymbol{\zeta}'} \end{bmatrix} \\ &= \left[\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix}\right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix}\right)}\right] \mathcal{C} \\ &\quad + \left[\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\boldsymbol{\zeta}^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}\right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\boldsymbol{\zeta}^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix}\right)}\right]. \end{aligned}$$

This means that our estimators may be expressed as

$$\begin{aligned}\widehat{\mathcal{C}} &= \left(\sum_{j=0}^{T-1} \dot{\mathbf{Z}}_j^{\mathbf{f}} \dot{\mathbf{Z}}_j^{\mathbf{f}^*}\right)^{-1} \sum_{j=0}^{T-1} \dot{\mathbf{Z}}_j^{\mathbf{f}} \breve{\mathbf{Z}}_j^{\mathbf{x}^*} \\ &= \mathcal{C} + \left\{ \frac{1}{T} \sum_{j=0}^{T-1} \left[\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right)} \right] \right\}^{-1} \\ &\quad \times \frac{1}{T} \sum_{j=0}^{T-1} \left[\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\zeta^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\zeta^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right)} \right].\end{aligned}$$

Next, we consider the expectation of $r \times 1$ vector $\mathbf{z}_j^{\mathbf{f}} z_j^{\zeta_i^*}$ for a given i and j.

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\theta}} \left[\mathbf{z}_j^{\mathbf{f}} z_j^{\zeta_i^*} \right] &= \mathsf{E}_{\boldsymbol{\theta}} \left[\mathbf{z}_j^{\mathbf{f}} \mathsf{E}_{\boldsymbol{\theta}} [\overline{z_j^{\zeta_i}} | \mathbf{z}_j^{\mathbf{f}}] \right] \\ &\qquad \text{(by the Law of Iterated Expectations)} \\ &= \mathbf{0}_{r \times 1} \quad \text{(by Assumption A2).}\end{aligned}$$

It follows that $\mathsf{E}_{\boldsymbol{\theta}}[\widehat{\mathcal{C}}] = \mathcal{C}$ and we have unbiasedness.

As regards consistency, we notice first the matrices involved in the inverse term are Hermitian and non-negative definite so that

$$\begin{aligned}&\left\| \frac{1}{T} \sum_{j=0}^{T-1} \left[\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right)} \right] \right\| \\ &\geq \left\| \frac{1}{T} \sum_{j=0}^{T-1} \left[\mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \otimes \begin{pmatrix} 1 & e^{i2\pi\omega_j} \\ e^{-i2\pi\omega_j} & 1 \end{pmatrix} \right] \right\| \\ &= \left\| \frac{1}{T} \sum_{j=0}^{T-1} \mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} \right\|.\end{aligned}$$

We note that

$$\begin{aligned}\frac{1}{T} \sum_{j=0}^{T-1} \mathbf{z}_j^{\mathbf{f}} \mathbf{z}_j^{\mathbf{f}^*} &= \frac{1}{T^2} \sum_{j=0}^{T-1} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbf{f}_t \mathbf{f}_s' e^{i2\pi\omega_j(t-s)} = \frac{1}{T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbf{f}_t \mathbf{f}_s' \sum_{j=0}^{T-1} e^{i2\pi\omega_j(t-s)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t' \\ &= \Gamma_{\mathbf{f}} + O_p(1/\sqrt{T}),\end{aligned}$$

due to ergodicity of $\mathbf{f}_t$, in turn a consequence of Assumption A1. In other words, the norm of the inverse term is bounded away from zero by a positive term. Moreover, this result is independent of i for $i = 1, ..., N$.

Next, we consider the error term,

$$\begin{aligned}
&\left\| \frac{1}{T} \sum_{j=0}^{T-1} \left[\left(\mathbf{z}_j^{\mathbf{f}} z_j^{\zeta_i^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right) + \overline{\left(\mathbf{z}_j^{\mathbf{f}} z_j^{\zeta_i^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right)} \right] \right\| \\
&\leq 2 \left\| \frac{1}{T} \sum_{j=0}^{T-1} \left(\mathbf{z}_j^{\mathbf{f}} z_j^{\zeta_i^*} \otimes \begin{pmatrix} 1 \\ e^{-i2\pi\omega_j} \end{pmatrix} \right) \right\|.
\end{aligned}$$

Looking at the squared expectation of the necessary expression, we observe that there exists a finite positive constant M^{Prop4}, independent of observations i for $i = 1, ..., N$, such that

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\theta}} \left[T \left\| \frac{1}{T} \sum_{j=0}^{T-1} \mathbf{z}_j^{\mathbf{f}} \overline{z_j^{\zeta_i}} \right\|^2 \right] &= \frac{1}{T} \mathsf{E}_{\boldsymbol{\theta}} \left[\sum_{j=0}^{T-1} \left(\mathbf{z}_j^{\mathbf{f}} \overline{z_j^{\zeta_i}} \right)^* \sum_{j'=0}^{T-1} \left(\mathbf{z}_{j'}^{\mathbf{f}} \overline{z_{j'}^{\zeta_i}} \right) \right] \\
&= \frac{1}{T} \sum_{k=1}^{r} \sum_{j=0}^{T-1} \sum_{j'=0}^{T-1} \mathsf{E}_{\boldsymbol{\theta}} \left[z_j^{f_k} \overline{z_{j'}^{f_k}} \right] \mathsf{E}_{\boldsymbol{\theta}} \left[z_j^{\zeta_i} \overline{z_{j'}^{\zeta_i}} \right] \\
&= \frac{1}{T} \sum_{k=1}^{r} \sum_{j=0}^{T-1} \mathsf{E}_{\boldsymbol{\theta}} \left[z_j^{f_k} \overline{z_j^{f_k}} \right] \mathsf{E}_{\boldsymbol{\theta}} \left[z_j^{\zeta_i} \overline{z_j^{\zeta_i}} \right] \\
&\leq M^{Prop4},
\end{aligned}$$

due to Assumptions A2 and A6. The result on $\sqrt{T}$-consistency of the estimator follows from Chebyshev's inequality and an application of the continuous mapping theorem. Moreover, since M^{Prop4} is independent of i for $i = 1, ..., N$, the consistency result is uniform with respect to the observations. This completes the proof of Proposition 2.4.

2.10.6 Proof of Proposition 2.5

In this section, we prove N, T consistency of our initialisation method for the factors.[27] Let us begin by recalling the definition of our estimator for the DFT of the factors:

$$\mathbf{z}_j^{\mathbf{f}^P} = \tilde{P}^*_{\mathbf{x}}(w_j)\mathbf{z}_j^{\mathbf{x}}/\sqrt{N}.$$

Step 1: Decomposition of $\mathbf{z}_j^{\mathbf{f}^P}$

Given the DFT of the observables equation of our true model,

$$\mathbf{z}_j^{\mathbf{x}} = C(e^{-i2\pi\omega_j})\mathbf{z}_j^{\mathbf{f}} + \mathbf{z}_j^{\zeta},$$

we have that

$$\begin{aligned}
\mathbf{z}_j^{\mathbf{f}^P} - \mathbf{z}_j^{\mathbf{f}} &= N^{-1/2}\tilde{P}^*_{\mathbf{x}}(w_j)\mathbf{z}_j^{\mathbf{x}} - \mathbf{z}_j^{\mathbf{f}} \\
&= N^{-1/2}\tilde{P}^*_{\mathbf{x}}(w_j)[C(e^{-i2\pi\omega_j})\mathbf{z}_j^{\mathbf{f}} + \mathbf{z}_j^{\zeta}] - \mathbf{z}_j^{\mathbf{f}} \\
&= [N^{-1/2}\tilde{P}^*_{\mathbf{x}}(w_j)C(e^{-i2\pi\omega_j}) - I_r]\mathbf{z}_j^{\mathbf{f}} + N^{-1/2}\tilde{P}^*_{\mathbf{x}}(w_j)\mathbf{z}_j^{\zeta} \\
&= [N^{-1/2}\tilde{P}^*_{\mathbf{x}}(w_j)N^{1/2}P_{\boldsymbol{\chi}}(w_j) - I_r]\mathbf{z}_j^{\mathbf{f}} + N^{-1/2}\tilde{P}^*_{\mathbf{x}}(w_j)\mathbf{z}_j^{\zeta} \\
&= [\tilde{P}^*_{\mathbf{x}}(w_j)P_{\boldsymbol{\chi}}(w_j) - I_r]\mathbf{z}_j^{\mathbf{f}} + N^{-1/2}\tilde{P}^*_{\mathbf{x}}(w_j)\mathbf{z}_j^{\zeta}.
\end{aligned}$$

Our goal will be to examine the asymptotics of the two terms in the above summation as N and T go to ∞.

Step 2: Intermediate Results

For convenience (in particular, to avoid over-burdening the notation needed below), let us assume just for the moment that $\kappa = 1$ which is true for the Bartlett lag-window. In this case, there is no need to explicitly carry around the additional rate associated with the bias in A9. For the purpose of the body of this proof, let us define $\rho_T = T/B_T \log B_T$ as in Forni et al. (2017, Appendix B. Proof of

[27] We combine techniques from various proofs in Forni et al. (2009), Doz et al. (2011) and Forni et al. (2017).

Proposition 7). We re-introduce the suppressed term in the last line of this proof.

Lemma 2.10.4. *For frequencies* $w_j = j/(2B_T + 1)$ *where* $j = 0, ..., 2B_T$, *as* $N \to \infty$ *and* $T \to \infty$,

(i) $\max_j N^{-1}\|\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j)\| = O_p(\rho_T^{-1/2})$;

(ii) $\max_j N^{-1}\|\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\boldsymbol{\chi}}(w_j)\| = O_p(\max(N^{-1}, \rho_T^{-1/2}))$;

Proof. For statement (i), we note that

$$\begin{aligned}
\|\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j)\|^2 & \\
&= \lambda_1\{[\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j)][\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j)]^*\} \\
&\leq \mathrm{Tr}\left([\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j)][\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j)]^*\right) \\
&= \sum_{i=1}^{N}\sum_{s=1}^{N} |\tilde{g}_{\mathbf{x},is}^{B_T}(w_j) - g_{\mathbf{x},is}(w_j)|^2.
\end{aligned}$$

So

$$\begin{aligned}
E[\max_j N^{-2}\|\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j)\|^2] & \\
&\leq E\left[\max_j N^{-2}\sum_{i=1}^{N}\sum_{s=1}^{N} |\tilde{g}_{\mathbf{x},is}^{B_T}(w_j) - g_{\mathbf{x},is}(w_j)|^2\right] \\
&= N^{-2}\sum_{i=1}^{N}\sum_{s=1}^{N} E\left[\max_j |\tilde{g}_{\mathbf{x},is}^{B_T}(w_j) - g_{\mathbf{x},is}(w_j)|^2\right] \\
&\leq M_{11} \times \rho_T^{-1} \quad \text{(by Assumption A9)}.
\end{aligned}$$

The statement follows from an application of Chebyshev's inequality.

For statement (ii), the identity $G_{\mathbf{x}}(w_j) = G_{\boldsymbol{\chi}}(w_j) + G_{\zeta}(w_j)$ implies that $\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\boldsymbol{\chi}}(w_j) = \tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j) + G_{\zeta}(w_j)$. Then, by the triangle inequality for the matrix norm,

$$\|\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\boldsymbol{\chi}}(w_j)\| \leq \|\tilde{G}_{\mathbf{x}}^{B_T}(w_j) - G_{\mathbf{x}}(w_j)\| + \|G_{\zeta}(w_j)\|.$$

Statement (ii) then follows from statement (i) and the fact that $\|G_\zeta(w_j)\| = |\lambda_1\{G_\zeta(w_j)\}| = O(1)$ by Assumption A6. □

Lemma 2.10.5. *For frequencies $w_j = j/(2B_T + 1)$ where $j = 0, ..., 2B_T$, as $N \to \infty$ and $T \to \infty$,*

(i) $\max_j N^{-1}|\lambda_k\{\tilde{G}^{B_T}_{\mathbf{x}}(w_j)\} - \lambda_k\{G_{\boldsymbol{\chi}}(w_j)\}| = O_p(\max(N^{-1}, \rho_T^{-1/2}))$ *for* $k = 1, ..., r$;

(ii) $\max_j \|\Lambda_{\boldsymbol{\chi}}(w_j)/N\|$ *and* $\max_j \|(\Lambda_{\boldsymbol{\chi}}(w_j)/N)^{-1}\|$, *which depend on* N, *are* $O(1)$;

(iii) $\max_j \|\tilde{\Lambda}_{\mathbf{x}}(w_j)/N\|$ *and* $\max_j \|(\tilde{\Lambda}_{\mathbf{x}}(w_j)/N)^{-1}\|$, *which depend on* N *and* T, *are* $O_p(1)$.

Proof. We use the consequence of Weyl's inequality that for any two $r \times r$ Hermitian matrices A and B,

$$|\lambda_k\{A + B\} - \lambda_k\{A\}| \leq \|B\|, \quad \text{for } k = 1, ..., r.$$

Setting $A = G_{\boldsymbol{\chi}}(w_j)$ and $B = \tilde{G}^{B_T}_{\mathbf{x}}(w_j) - G_{\boldsymbol{\chi}}(w_j)$ we have that for $k = 1, ..., r$,

$$|\lambda_k\{\tilde{G}^{B_T}_{\mathbf{x}}(w_j)\} - \lambda_k\{G_{\boldsymbol{\chi}}(w_j)\}| \leq \|\tilde{G}^{B_T}_{\mathbf{x}}(w_j) - G_{\boldsymbol{\chi}}(w_j)\|$$

Statement (i) follows from Lemma 2.10.4(ii).

Statement (ii) is proved by noting that $\|\Lambda_{\boldsymbol{\chi}}(w_j)/N\| = |\lambda_1\{G_{\boldsymbol{\chi}}(w_j)\}|/N$ is bounded in probability uniformly with respect to the frequency due to Assumption A3. The same applies to $\|(\Lambda_{\boldsymbol{\chi}}(w_j)/N)^{-1}\| = N/|\lambda_r\{G_{\boldsymbol{\chi}}(w_j)\}|$.

Statement (iii) follows from statement (i). □

Lemma 2.10.6. *For frequencies* $w_j = j/(2B_T + 1)$ *where* $j = 0, ..., 2B_T$, *as* $N \to \infty$ *and* $T \to \infty$,

(i) $\max_j N^{-1} \|P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)\tilde{\Lambda}_{\mathbf{x}}(w_j) - \Lambda_{\boldsymbol{\chi}}(w_j)P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)\|$
$= O_p(\max(N^{-1}, \rho_T^{-1/2}))$;

(ii) $\max_j \|\tilde{P}^*_{\mathbf{x}}(w_j)P_{\boldsymbol{\chi}}(w_j)P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j) - I_r\| = O_p(\max(N^{-1}, \rho_T^{-1/2}))$;

(iii) there exist complex diagonal orthogonal matrices, depending on N *and* T, *denoted by* $\tilde{Q}_r(w_j) = \mathrm{diag}(\tilde{q}_{1,j}, \tilde{q}_{2,j}, ..., \tilde{q}_{r,j})$ *with* $|\tilde{q}_{k,j}|^2 = 1$ *for* $k = 1, ..., r$ *such that*

$$\max_j \|\tilde{P}^*_{\mathbf{x}}(w_j)P_{\boldsymbol{\chi}}(w_j) - \tilde{Q}_r(w_j)\| = O_p(\max(N^{-1}, \rho_T^{-1/2})).$$

Proof. Since $\tilde{P}_{\mathbf{x}}(w_j)\tilde{\Lambda}_{\mathbf{x}}(w_j) = \tilde{G}^{B_T}_{\mathbf{x}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)$ and $\Lambda_{\boldsymbol{\chi}}(w_j)P^*_{\boldsymbol{\chi}}(w_j) = P^*_{\boldsymbol{\chi}}(w_j)G_{\boldsymbol{\chi}}(w_j)$, we have

$$\begin{aligned}
&\|P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)(\tilde{\Lambda}_{\mathbf{x}}(w_j)/N) - (\Lambda_{\boldsymbol{\chi}}(w_j)/N)P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)\| \\
&\qquad = N^{-1}\|P^*_{\boldsymbol{\chi}}(w_j)\tilde{G}^{B_T}_{\mathbf{x}}(w_j)\tilde{P}_{\mathbf{x}}(w_j) - P^*_{\boldsymbol{\chi}}(w_j)G_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)\| \\
&\qquad = N^{-1}\|P^*_{\boldsymbol{\chi}}(w_j)[\tilde{G}^{B_T}_{\mathbf{x}}(w_j) - G_{\boldsymbol{\chi}}(w_j)]\tilde{P}_{\mathbf{x}}(w_j)\| \\
&\qquad \leq N^{-1}\|P^*_{\boldsymbol{\chi}}(w_j)\|\|\tilde{G}^{B_T}_{\mathbf{x}}(w_j) - G_{\boldsymbol{\chi}}(w_j)\|\|\tilde{P}_{\mathbf{x}}(w_j)\| \\
&\qquad\qquad \text{(by submultiplicativity of the norm)} \\
&\qquad = N^{-1}\|\tilde{G}^{B_T}_{\mathbf{x}}(w_j) - G_{\boldsymbol{\chi}}(w_j)\|.
\end{aligned}$$

Statement (i) follows from Lemma 2.10.4(ii).

For statement (ii), we define

$$
\begin{aligned}
A &= \tilde{P}^*_{\mathbf{x}}(w_j)P_{\boldsymbol{\chi}}(w_j)P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j) \\
&= \tilde{P}^*_{\mathbf{x}}(w_j)P_{\boldsymbol{\chi}}(w_j)P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)\left(\frac{\tilde{\Lambda}_{\mathbf{x}}(w_j)}{N}\right)\left(\frac{\tilde{\Lambda}_{\mathbf{x}}(w_j)}{N}\right)^{-1}, \\
B &= \tilde{P}^*_{\mathbf{x}}(w_j)P_{\boldsymbol{\chi}}(w_j)\left(\frac{\Lambda_{\boldsymbol{\chi}}(w_j)}{N}\right)P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)\left(\frac{\tilde{\Lambda}_{\mathbf{x}}(w_j)}{N}\right)^{-1} \\
&= \frac{1}{N}\tilde{P}^*_{\mathbf{x}}(w_j)G_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)\left(\frac{\tilde{\Lambda}_{\mathbf{x}}(w_j)}{N}\right)^{-1}, \\
C &= \frac{1}{N}\tilde{P}^*_{\mathbf{x}}(w_j)\tilde{G}^{B_T}_{\mathbf{x}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)\left(\frac{\tilde{\Lambda}_{\mathbf{x}}(w_j)}{N}\right)^{-1} \\
&= \left(\frac{\tilde{\Lambda}_{\mathbf{x}}(w_j)}{N}\right)\left(\frac{\tilde{\Lambda}_{\mathbf{x}}(w_j)}{N}\right)^{-1} \\
&= I_r
\end{aligned}
$$

and note that $\|A - C\| \leq \|A - B\| + \|B - C\|$. Both terms on the right are bounded in probability uniformly with respect to the frequency with rate $O_p(\max(N^{-1}, \rho_T^{-1/2}))$, the first due to statement (i) along with Lemma 2.10.5(iii) and the second due to Lemma 2.10.4(ii) along with Lemma 2.10.5(iii).

For statement (iii), we consider a single element of the matrix on the left side of statement (i). Denoting by $P_{\boldsymbol{\chi},k}(w_j)$ and $\tilde{P}_{\mathbf{x},k}(w_j)$ the k^{th} columns of $P_{\boldsymbol{\chi}}(w_j)$ and $\tilde{P}_{\mathbf{x}}(w_j)$ respectively, we have from statement (i) that

$$
\max_j N^{-1}|(\lambda_l\{\tilde{G}^{B_T}_{\mathbf{x}}(w_j)\} - \lambda_k\{G_{\boldsymbol{\chi}}(w_j)\})P^*_{\boldsymbol{\chi},k}(w_j)\tilde{P}_{\mathbf{x},l}(w_j)| = O_p(\max(N^{-1}, \rho_T^{-1/2}))
$$

for $k, l = 1, ..., r$.

We recall the convergence result for estimated dynamic eigenvalues of $\boldsymbol{\chi}_t$ in Lemma 2.10.5(i). Then, due to Assumption A3, which ensures asymptotic separation of the dynamic eigenvalues, we have that for $k \neq l$, the expression $N^{-1}|(\lambda_l\{\tilde{G}^{B_T}_{\mathbf{x}}(w_j)\} - \lambda_k\{G_{\boldsymbol{\chi}}(w_j)\})$ is bounded away from 0 in probability uniformly with respect to the frequency. Therefore, it must hold that the modulus

of any off-diagonal term of $\tilde{P}^*_{\mathbf{x}}(w_j)P_{\boldsymbol{\chi}}(w_j)$ is $O_p(\max(N^{-1}, \rho_T^{-1/2}))$ uniformly with respect to the frequency.

As regards the diagonal terms, we consider a single element of the matrix on the left side of statement (ii). For $k = 1, ..., r$, we have

$$\begin{aligned}\max_j |\tilde{P}^*_{\mathbf{x},k}(w_j)P_{\boldsymbol{\chi}}(w_j)P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x},k}(w_j) - 1| \\ &= \max_j \left| \sum_{l=1}^{r} |\tilde{P}^*_{\mathbf{x},k}(w_j)P_{\boldsymbol{\chi},l}(w_j)|^2 - 1 \right| \\ &= O_p(\max(N^{-1}, \rho_T^{-1/2})).\end{aligned}$$

Consequently,

$$\begin{aligned}\max_j ||\tilde{P}^*_{\mathbf{x},k}(w_j)P_{\boldsymbol{\chi},k}(w_j)|^2 - 1| \\ &= \max_j |(|\tilde{P}^*_{\mathbf{x},k}(w_j)P_{\boldsymbol{\chi},k}(w_j)| + 1)(|\tilde{P}^*_{\mathbf{x},k}(w_j)P_{\boldsymbol{\chi},k}(w_j)| - 1)| \\ &= O_p(\max(N^{-1}, \rho_T^{-1/2})).\end{aligned}$$

Thus, the modulus of any diagonal term converges in probability to 1 uniformly with respect to the frequency at rate $O_p(\max(N^{-1}, \rho_T^{-1/2}))$.

Statement (iii) follows. □

Lemma 2.10.7. *For frequencies $w_j = j/(2B_T + 1)$ where $j = 0, ..., 2B_T$, as $N \to \infty$ and $T \to \infty$,*

$$\max_j \|P_{\boldsymbol{\chi}}(w_j)\tilde{Q}_r(w_j) - \tilde{P}_{\mathbf{x}}(w_j)\| = O_p(\max(N^{-1}, \rho_T^{-1/2})),$$

uniformly in i.

Proof. Since $P^*_{\boldsymbol{\chi}}(w_j)P_{\boldsymbol{\chi}}(w_j) = I_r$, we first note that $\|P_{\boldsymbol{\chi}}(w_j)\| = O(1)$. Next, we

define

$$\begin{aligned}
D &= P_{\boldsymbol{\chi}}(w_j)\tilde{Q}_r(w_j), \\
E &= P_{\boldsymbol{\chi}}(w_j)[P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)] \\
&= P_{\boldsymbol{\chi}}(w_j)[P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)(\tilde{\Lambda}_{\mathbf{x}}(w_j)/N)](\tilde{\Lambda}_{\mathbf{x}}(w_j)/N)^{-1}, \\
F &= P_{\boldsymbol{\chi}}(w_j)[(\Lambda_{\boldsymbol{\chi}}(w_j)/N)P^*_{\boldsymbol{\chi}}(w_j)\tilde{P}_{\mathbf{x}}(w_j)](\tilde{\Lambda}_{\mathbf{x}}(w_j)/N)^{-1} \\
&= [N^{-1}G_{\boldsymbol{\chi}}(w_j)]\tilde{P}_{\mathbf{x}}(w_j)(\tilde{\Lambda}_{\mathbf{x}}(w_j)/N)^{-1}, \\
G &= [N^{-1}\tilde{G}^{B_T}_{\mathbf{x}}(w_j)]\tilde{P}_{\mathbf{x}}(w_j)(\tilde{\Lambda}_{\mathbf{x}}(w_j)/N)^{-1} \\
&= \tilde{P}_{\mathbf{x}}(w_j)
\end{aligned}$$

and note that $\|D - G\| \leq \|D - E\| + \|E - F\| + \|F - G\|$. All terms on the right are $O_p(\max(N^{-1}, \rho_T^{-1/2}))$, the first because of Lemma 2.10.6(iii), the second because of Lemma 2.10.6(i) and Lemma 2.10.5(iii), and finally the third because of Lemma 2.10.4(ii) and Lemma 2.10.5(iii).

The result follows. □

Step 3: Asymptotic Analysis of $\mathbf{z}_j^{\mathbf{f}^P}$ as N and T go to ∞

We recall from Step 1 that we have

$$\mathbf{z}_j^{\mathbf{f}^P} - \mathbf{z}_j^{\mathbf{f}} = [\tilde{P}^*_{\mathbf{x}}(w_j)P_{\boldsymbol{\chi}}(w_j) - I_r]\mathbf{z}_j^{\mathbf{f}} + N^{-1/2}\tilde{P}^*_{\mathbf{x}}(w_j)\mathbf{z}_j^{\zeta}.$$

Dynamic eigenvectors are defined up to post-multiplication by a complex diagonal orthogonal matrix with unit-modulus diagonal elements. In particular, using the dynamic eigenvectors $\Pi_{\boldsymbol{\chi}}(w_j) = P_{\boldsymbol{\chi}}(w_j)\tilde{Q}_r(w_j)$, the result of Lemma 2.10.6(iii) could be expressed as

$$\max_j \|\tilde{P}^*_{\mathbf{x}}(w_j)\Pi_{\boldsymbol{\chi}}(w_j) - I_r\| = O_p(\max(N^{-1}, \rho_T^{-1/2})).$$

We thus conduct our subsequent analysis using the expression $[\mathbf{z}_j^{\mathbf{f}^P} - \tilde{Q}_r(w_j)\mathbf{z}_j^{\mathbf{f}}]$ rather than $[\mathbf{z}_j^{\mathbf{f}^P} - \mathbf{z}_j^{\mathbf{f}}]$. We have

$$\begin{aligned}\mathbf{z}_j^{\mathbf{f}^P} - \tilde{Q}_r(w_j)\mathbf{z}_j^{\mathbf{f}} &= [\tilde{P}_{\mathbf{x}}^*(w_j)P_{\boldsymbol{\chi}}(w_j) - \tilde{Q}_r(w_j)]\tilde{Q}_r(w_j)\mathbf{z}_j^{\mathbf{f}} + N^{-1/2}\tilde{P}_{\mathbf{x}}^*(w_j)\mathbf{z}_j^{\zeta} \\ &= [\tilde{P}_{\mathbf{x}}^*(w_j)\Pi_{\boldsymbol{\chi}}(w_j) - I_r]\mathbf{z}_j^{\mathbf{f}} + N^{-1/2}\tilde{P}_{\mathbf{x}}^*(w_j)\mathbf{z}_j^{\zeta}.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathbf{z}_j^{\mathbf{f}^P} - \tilde{Q}_r(w_j)\mathbf{z}_j^{\mathbf{f}}\| &= \|[\tilde{P}_{\mathbf{x}}^*(w_j)\Pi_{\boldsymbol{\chi}}(w_j) - I_r]\mathbf{z}_j^{\mathbf{f}} + N^{-1/2}\tilde{P}_{\mathbf{x}}^*(w_j)\mathbf{z}_j^{\zeta}\| \\ &\leq \|[\tilde{P}_{\mathbf{x}}^*(w_j)\Pi_{\boldsymbol{\chi}}(w_j) - I_r]\mathbf{z}_j^{\mathbf{f}}\| + \|N^{-1/2}\tilde{P}_{\mathbf{x}}^*(w_j)\mathbf{z}_j^{\zeta}\| \\ &\qquad \text{(by the triangle inequality for the matrix norm)} \\ &\leq \|[\tilde{P}_{\mathbf{x}}^*(w_j)\Pi_{\boldsymbol{\chi}}(w_j) - I_r]\|\|\mathbf{z}_j^{\mathbf{f}}\| + \|N^{-1/2}\tilde{P}_{\mathbf{x}}^*(w_j)\mathbf{z}_j^{\zeta}\| \\ &\qquad \text{(by submultiplicativity of the norm).}\end{aligned}$$

Since $\max_j \|\mathbf{z}_j^{\mathbf{f}}\| = O_p(1)$ by Assumption A2, the first term above converges in probability to 0 uniformly with respect to the frequency at rate $O_p(\max(N^{-1}, \rho_T^{-1/2}))$ due to Lemma 2.10.6(iii). It remains to analyse the second term.

Now,

$$\begin{aligned}\|N^{-1/2}\tilde{P}_{\mathbf{x}}^*(w_j)\mathbf{z}_j^{\zeta}\| &= N^{-1/2}\|[\tilde{P}_{\mathbf{x}}^*(w_j) - \tilde{Q}_r^*(w_j)P_{\boldsymbol{\chi}}^*(w_j) + \tilde{Q}_r^*(w_j)P_{\boldsymbol{\chi}}^*(w_j)]\mathbf{z}_j^{\zeta}\| \\ &= N^{-1/2}\|[\tilde{P}_{\mathbf{x}}^*(w_j) - \Pi_{\boldsymbol{\chi}}^*(w_j)]\mathbf{z}_j^{\zeta} + \Pi_{\boldsymbol{\chi}}^*(w_j)\mathbf{z}_j^{\zeta}\| \\ &\leq \|N^{-1/2}[\tilde{P}_{\mathbf{x}}^*(w_j) - \Pi_{\boldsymbol{\chi}}^*(w_j)]\mathbf{z}_j^{\zeta}\| + N^{-1/2}\|\Pi_{\boldsymbol{\chi}}^*(w_j)\mathbf{z}_j^{\zeta}\| \\ &\leq \|N^{-1/2}[\tilde{P}_{\mathbf{x}}^*(w_j) - \Pi_{\boldsymbol{\chi}}^*(w_j)]\|\|\mathbf{z}_j^{\zeta}\| + N^{-1/2}\|\Pi_{\boldsymbol{\chi}}^*(w_j)\mathbf{z}_j^{\zeta}\|.\end{aligned}$$

The first term on the right is $O_p(N^{-1/2}\max(N^{-1}, \rho_T^{-1/2}))$ due to Lemma 2.10.7. As noted above, $\|\mathbf{z}_j^{\mathbf{f}}\| = O_p(1)$. Further, the final term on the right is $O_p(N^{-1/2})$

since

$$\begin{aligned} E[\|\Pi^*_{\boldsymbol{\chi}}(w_j)\mathbf{z}_j^{\zeta}\|^2] &= E[\lambda_1\{\mathbf{z}_j^{\zeta^*}\Pi_{\boldsymbol{\chi}}(w_j)\Pi^*_{\boldsymbol{\chi}}(w_j)\mathbf{z}_j^{\zeta}\}] \\ &= E[\mathbf{z}_j^{\zeta^*}\Pi_{\boldsymbol{\chi}}(w_j)\Pi^*_{\boldsymbol{\chi}}(w_j)\mathbf{z}_j^{\zeta}] \\ &= E[\mathrm{Tr}(\mathbf{z}_j^{\zeta^*}\Pi_{\boldsymbol{\chi}}(w_j)\Pi^*_{\boldsymbol{\chi}}(w_j)\mathbf{z}_j^{\zeta})] \\ &= \mathrm{Tr}(\Pi^*_{\boldsymbol{\chi}}(w_j)G_{\zeta}(w_j)\Pi_{\boldsymbol{\chi}}(w_j)) \\ &\leq r\lambda_1\{G_{\zeta}(w_j)\} = O(1) \\ &\qquad \text{(by Assumption A4).} \end{aligned}$$

The result follows that $\max_j \|\mathbf{z}_j^{\mathbf{f}^P} - \tilde{Q}_r(w_j)\mathbf{z}_j^{\mathbf{f}}\| = O_p(\max(N^{-1/2}, \rho_T^{-1/2}))$. Finally, by relaxing the assumption that $\kappa = 1$ and redoing the entire proof while carrying through the additional bias term everywhere would yield the required result of Proposition 2.5.

2.10.7 Proof of Proposition 2.6

Proof. We first note that if we were to replace all the summations over T points in the proof of Proposition 2.4 by summations over $2B_T + 1$ points, we would obtain a rate of $\sqrt{B_T}$ instead of $\sqrt{T}$. Combining this result with that of Proposition 2.5, we recognise that our error is

$$O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \frac{1}{\sqrt{B_T^{2\kappa}}}, \sqrt{\frac{B_T \log B_T}{T}}\right)\right).$$

Finally, noticing that the term $1/\sqrt{B_T}$ dominates $1/\sqrt{B_T^{2\kappa}}$ whenever $\kappa > 1/2$, the required result follows. We explicitly note here that for all commonly used lag windows, indeed, κ is typically 1 or 2; see (Priestley, 1982, p. 463). □

2.10.8 Proof of Proposition 2.7

Proof. We organise our proof into two parts. The first part, which we refer to as a technical remark, explains why we can claim that there exists a point, say $\hat{\boldsymbol{\theta}}^{\dagger\dagger}$, to which the spectral EM iterates $\{\hat{\boldsymbol{\theta}}^{(k)}\}$ converge (as $k \to \infty$); and further, that this point is a local maximum. This property can easily be seen for the general case in Wu (1983) (which we have summarised in considerable detail in Appendix 2.11) and for the time domain analogue of our specific setting in Barigozzi and Luciani (2022, Lemma 21). The logic in the frequency domain is identical.

Part 1. Technical remark:

(i) Is the problem well-defined?

Consider our d-dimensional vector of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (see Section 2.3), for which parameter space $\boldsymbol{\Theta}$ is such that $\boldsymbol{\Theta} \subset \mathbb{R}^d$ where $d = 2rN + N + r^2 + r$. For finite dimensional parameter spaces, the Heine-Borel theorem[28] provides a simple characterisation of which sets are compact, but for potentially infinite dimensional parameter spaces, the situation is more complicated. Nevertheless, in our setting, due to the form of the (quasi-) log likelihood (see Section 2.5.2), the M step is greatly simplified. Indeed, for any iteration $k \geq 0$, the M step requires solving the N individual maximisations

$$\left(\hat{\boldsymbol{c}}_i^{(k+1)'}, \hat{\Gamma}_{\xi_i}^{(k+1)}\right)' = \underset{(\boldsymbol{c}_i', \Gamma_{\xi_i})' \in \mathcal{O}_1}{\arg\max} \, \mathsf{E}_{\hat{\boldsymbol{\theta}}^{(k)}} \left[l_i \left(x_{i1}, ..., x_{iT} | F; \boldsymbol{c}_i, \Gamma_{\xi_i} \right) | X \right]$$

for $i = 1, ..., N$, where (i) the function l_i refers to the marginal contribution of the i-th observation to the overall Whittle log-likelihood computed with respect to the coarse grid of $2B_T + 1$ frequencies (see Sections 2.5.2 and 2.7.3 for details); and (ii) $\mathcal{O}_1 = [-m_c, m_c]^{2r} \times [m_\xi{}^{-1}, m_\xi] \subset \mathbb{R}^{2r+1}$, with "$\times$" denoting the Cartesian

[28] If a set $\mathcal{S}$ of real numbers is closed and bounded, then the set $\mathcal{S}$ is compact. That is, if a set $\mathcal{S}$ of real numbers is closed and bounded, then every open cover of the set $\mathcal{S}$ has a finite subcover.

product; and the finite dimensional maximisation

$$\left(\text{vech}(\hat{A}^{(k+1)}), \hat{\gamma}_1^{(k+1)}, ..., \hat{\gamma}_r^{(k+1)}\right)' = \underset{(\text{vech}(A),\gamma_1,...,\gamma_r)' \in \mathcal{O}_2}{\arg\max} \mathsf{E}_{\hat{\boldsymbol{\theta}}^{(k)}} \left[l\left(F|A, \gamma_1, ..., \gamma_r\right) |X \right],$$

where (i) A is the matrix of autoregressive coefficients such that $A(L) = I_r - AL$; and (ii) $\mathcal{O}_2 = [-m_a, m_a]^{r^2} \times [{m_\gamma}^{-1}, m_\gamma]^r \subset \mathbb{R}^{r^2+r}$. In this setup, it holds that $\boldsymbol{\Theta} = \mathcal{O}_1^N \times \mathcal{O}_2$. Recall further that we have Gaussianity due to Assumption A7.

Then for any iteration $k \geq 0$, the M step estimator given by

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\, \mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)})$$

is well-defined since $\mathcal{O}_1$ and $\mathcal{O}_2$ – each the product of finitely many compact sets – are compact sets, and $l\left(X, F; \boldsymbol{\theta}\right)$ is continuous and differentiable in the interior of each of those sets (Gourieroux and Monfort, 1995, Property 7.11, p.181–182). Further, it holds that $\hat{\boldsymbol{\theta}}^{(k+1)}$ is unique because the log likelihoods alluded to above are concave in their arguments.

(ii) Do we have convergence to a stationary point?

Next, by recalling that (i) $\mathcal{O}_1$ and $\mathcal{O}_2$ are compact sets, (ii) $l(X; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\Theta}$ and differentiable in the interior of $\boldsymbol{\Theta}$, and (iii) the point of initialisation is chosen such that $\hat{\boldsymbol{\theta}}^{(0)}$ is assumed to satisfy $l(X; \hat{\boldsymbol{\theta}}^{(0)}) > -\infty$, it follows that the sequence $\{l(X; \hat{\boldsymbol{\theta}}^{(k)})\}$ for $k \geq 0$ is bounded from above for any $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}$ (see Barigozzi and Luciani (2022, p.95) and Wu (1983, p.96–97)). Further, we recall the monotonicity property of the EM algorithm (which we prove in detail in Appendix 2.11.1), and we note that $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\phi})$ is continuous in both its arguments for arbitrary values $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. It follows by Wu (1983, Theorem 2) that we are guaranteed monotonic convergence of the likelihood:

$$\lim_{k \to \infty} l(X; \hat{\boldsymbol{\theta}}^{(k)}) = l(X; \hat{\boldsymbol{\theta}}^{\dagger\dagger}),$$

where $\hat{\boldsymbol{\theta}}^{\dagger\dagger}$ is a stationary point.

(iii) Is the stationary point a local maximum?

Given that $\mathcal{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)})$ has a unique maximum, and since $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\phi})$ is continuous in both its arguments for arbitrary values $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ and its gradient $\nabla_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\phi})$ is continuous in $\boldsymbol{\theta}$, it follows from Wu (1983, Theorem 3) that $\hat{\boldsymbol{\theta}}^{\dagger\dagger}$ is a local maximum. (See also Barigozzi and Luciani (2022, Proof of Lemma 21).)

(iv) Do EM iterates also converge?

Further, we know by Wu (1983, Condition 1) that, under Gaussianity, there exists a real positive constant, say $M_{\mathcal{Q}}$, such that

$$\mathcal{Q}(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) \geq M_{\mathcal{Q}} \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\|^2$$

for all k. Together with monotonicity of the likelihood and the result under point *(iii)* above, the latter constitutes a sufficient condition for:

$$\lim_{k \to \infty} \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\| = 0,$$

Further details can be seen in Wu (1983, p.100). Finally, it follows from Wu (1983, Theorem 6) that

$$\lim_{k \to \infty} \boldsymbol{\theta}^{(k)} = \hat{\boldsymbol{\theta}}^{\dagger\dagger}.$$

Below, we will focus just on $\hat{\boldsymbol{c}}_i^{\dagger\dagger}$, a subset of the vector $\hat{\boldsymbol{\theta}}^{\dagger\dagger}$.

Part 2. Main proof:

We have

$$\left\| \hat{\boldsymbol{c}}_i^{SpEM} - \boldsymbol{c}_i \right\| \leq \left\| \hat{\boldsymbol{c}}_i^{SpEM} - \hat{\boldsymbol{c}}_i^{\dagger\dagger} \right\| + \left\| \hat{\boldsymbol{c}}_i^{\dagger\dagger} - \hat{\boldsymbol{c}}_i^{\dagger} \right\| + \left\| \hat{\boldsymbol{c}}_i^{\dagger} - \boldsymbol{c}_i \right\|,$$

where $\hat{\boldsymbol{c}}_i^{\dagger\dagger}$ denotes the point to which the Spectral EM algorithm converges as the number of iterations go to infinity,

$$\hat{\boldsymbol{c}}_i^{\dagger\dagger} = \lim_{k \to \infty} \hat{\boldsymbol{c}}_i^{(k)},$$

and $\hat{\boldsymbol{c}}_i^{\dagger}$ denotes the global maximiser of $l(X;\boldsymbol{\theta})$ for $i = 1, ..., N$.

The third term on the right-hand side is simply the error of the quasi-maximum likelihood estimator where the Whittle likelihood is computed via summations over $2B_T + 1$ points. By Proposition 2.4 and Remark 3, as $N, T \to \infty$,

$$\left\| \hat{\boldsymbol{c}}_i^{\dagger} - \boldsymbol{c}_i \right\| = O_p\left(\max\left(\frac{1}{N}, \frac{1}{\sqrt{B_T}} \right) \right),$$

uniformly in i for $i = 1, ..., N$. See Barigozzi and Luciani (2022, Lemma 14(i)) and Bai and Li (2016, Theorem 1).

Similarly, the second term on the right-hand side is

$$\left\| \hat{\boldsymbol{c}}_i^{\dagger\dagger} - \hat{\boldsymbol{c}}_i^{\dagger} \right\| = O_p\left(\max\left(\frac{1}{N}, \frac{1}{\sqrt{B_T}} \right) \right),$$

uniformly in i for $i = 1, ..., N$, as implied by Barigozzi and Luciani (2022, Lemma 22(i)).

Now we consider the first term on the right-hand side. This represents the discrepancy between the Spectral EM estimator and the point to which the spectral EM algorithm converges as $k \to \infty$, a local maximiser of $l(X;\boldsymbol{\theta})$. Using key parts of Barigozzi and Luciani (2022, Proof of Lemma 23), we know there exists a finite positive real constant M_{P7} such that

$$\left\| \hat{\boldsymbol{c}}_i^{SpEM} - \hat{\boldsymbol{c}}_i^{\dagger\dagger} \right\| \leq \left\| \hat{\boldsymbol{c}}_i^{(0)} - \hat{\boldsymbol{c}}_i^{\dagger\dagger} \right\| \left\| \boldsymbol{R}\left(\hat{\boldsymbol{c}}_i^{\dagger\dagger} \right) \right\|^{k^{SpEM}} + M_{P7} k^{SpEM} \left\| \hat{\boldsymbol{c}}_i^{(0)} - \hat{\boldsymbol{c}}_i^{\dagger\dagger} \right\|^2,$$

where $\boldsymbol{R}\left(\hat{\boldsymbol{c}}_i^{\dagger\dagger} \right)$ represents the relevant sub-matrix of a ratio (in the matrix sense) of conditional Hessians associated with the Whittle log-likelihood computed at the value of the local maximiser. It is sufficient for the purpose of this study to consider $\| \boldsymbol{R}\left(\hat{\boldsymbol{c}}_i^{\dagger\dagger} \right) \|$ as an $O_p(1)$ term.[29]

[29] In reality, it is vanishing at a rate dominated by the other rates in our procedure. See Barigozzi and Luciani (2022, Proof of Lemma 23) for further details.

Finally, by Proposition 2.6 and the results above, we have

$$\begin{aligned}
\left\|\hat{\boldsymbol{c}}_i^{(0)} - \hat{\boldsymbol{c}}_i^{\dagger\dagger}\right\| &\leq \left\|\hat{\boldsymbol{c}}_i^{(0)} - \boldsymbol{c}_i\right\| + \left\|\boldsymbol{c}_i - \hat{\boldsymbol{c}}_i^{\dagger}\right\| + \left\|\hat{\boldsymbol{c}}_i^{\dagger} - \hat{\boldsymbol{c}}_i^{\dagger\dagger}\right\| \\
&= O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \sqrt{\frac{B_T \log B_T}{T}}\right)\right) \\
&\quad + O_p\left(\max\left(\frac{1}{N}, \frac{1}{\sqrt{B_T}}\right)\right) \\
&\quad + O_p\left(\max\left(\frac{1}{N}, \frac{1}{\sqrt{B_T}}\right)\right) \\
&= O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \sqrt{\frac{B_T \log B_T}{T}}\right)\right),
\end{aligned}$$

and the required result follows. □

2.10.9 Proof of Proposition 2.8

Proof. The following algebraic identity will be useful. For arbitrary scalars a_0, b_0, a_1, and b_1,

$$\begin{aligned} a_1b_1 - a_0b_0 &= a_1b_1 - a_0b_0 + (a_1b_0 - a_1b_0) + (a_0b_1 - a_0b_1) + (a_0b_0 - a_0b_0) \\ &= a_1b_0 - a_0b_0 + a_0b_1 - a_0b_0 + a_1b_1 - a_1b_0 - a_0b_1 + a_0b_0 \\ &= (a_1 - a_0)b_0 + (b_1 - b_0)a_0 + (a_1 - a_0)(b_1 - b_0). \end{aligned}$$

Using similar reasoning,

$$\begin{aligned} \hat{z}_{j_q}^{\chi_i^{SpEM}} - z_{j_q}^{\chi_i} = &\left[\hat{\boldsymbol{c}}_{i,\cdot}^{SpEM}(e^{-i2\pi w_{j_q}}) - \boldsymbol{c}_{i,\cdot}(e^{-i2\pi w_{j_q}})\right]' \mathbf{z}_{j_q}^{\mathbf{f}} + \boldsymbol{c}'_{i,\cdot}(e^{-i2\pi w_{j_q}}) \left[\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{SpEM}} - \mathbf{z}_{j_q}^{\mathbf{f}}\right] \\ &+ \left[\hat{\boldsymbol{c}}_{i,\cdot}^{SpEM}(e^{-i2\pi w_{j_q}}) - \boldsymbol{c}_{i,\cdot}(e^{-i2\pi w_{j_q}})\right]' \left[\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{SpEM}} - \mathbf{z}_{j_q}^{\mathbf{f}}\right], \end{aligned}$$

so that

$$\begin{aligned} \left|\hat{z}_{j_q}^{\chi_i^{SpEM}} - z_{j_q}^{\chi_i}\right| \leq &\left|\left[\hat{\boldsymbol{c}}_{i,\cdot}^{SpEM}(e^{-i2\pi w_{j_q}}) - \boldsymbol{c}_{i,\cdot}(e^{-i2\pi w_{j_q}})\right]' \mathbf{z}_{j_q}^{\mathbf{f}}\right| + \left|\boldsymbol{c}'_{i,\cdot}(e^{-i2\pi w_{j_q}}) \left[\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{SpEM}} - \mathbf{z}_{j_q}^{\mathbf{f}}\right]\right| \\ &+ \left|\left[\hat{\boldsymbol{c}}_{i,\cdot}^{SpEM}(e^{-i2\pi w_{j_q}}) - \boldsymbol{c}_{i,\cdot}(e^{-i2\pi w_{j_q}})\right]' \left[\hat{\mathbf{z}}_{j_q}^{\mathbf{f}^{SpEM}} - \mathbf{z}_{j_q}^{\mathbf{f}}\right]\right|, \end{aligned}$$

for $q = 1, ..., 2B_T + 1$.

The first term on the right-hand side is

$$O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \sqrt{\frac{B_T \log B_T}{T}}\right)\right)$$

uniformly in $1 \leq i \leq N$ and $1 \leq q \leq 2B_T + 1$ by Proposition 2.7 and the fact that $\|\mathbf{z}_{j_q}^{\mathbf{f}}\| = O_p(1)$ uniformly in q as a consequence of weak stationarity of $\mathbf{f}_t$ emanating from Assumption A2.

The second term on the right-hand side is also

$$O_p\left(\max\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{B_T}}, \sqrt{\frac{B_T \log B_T}{T}}\right)\right)$$

uniformly in $1 \leq i \leq N$ and $1 \leq q \leq 2B_T + 1$ due to Theorem 2.1, Proposition 2.7 and the fact that $\|\boldsymbol{c}_{i,\cdot}(e^{-i2\pi w_{j_q}})\| = O(1)$ uniformly in $1 \leq i \leq N$ due to Assumption A2.

The third term on the right-hand side will converge in probability to zero as a product of the above two rates and thus be dominated by the first and second terms. The required result follows.

□

2.10.10 Proof of Proposition 2.9

Proof. The proof is analogous to the proof of Theorem 2.1, except where the factors therein are replaced with common components. □

2.10.11 Showing Assumption A9

The overall focus of our work is not strictly on consistent estimation of the spectral density *per se*; nevertheless, Assumption A9 is important for our study and its statement cannot be regarded as trivial. For this reason, we aim in this section to outline for the interested reader the steps that would be needed in order to derive the statement of Assumption A9 based on more primitive conditions.

Step 1. Defining Physical Dependence

The physical dependence measure espoused by Wu and Zaffaroni (2018) is characterised in terms of the L_p-norm of the difference between the observations and their constructed (or so-called "coupled") version using an independent and identically distributed copy of the innovation at time zero. A formal definition, following Forni et al. (2017), is given below.

Let $\boldsymbol{\epsilon}_t$ be an *iid* stochastic vector process, possibly infinite-dimensional, and let $z_t = F(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-1}, ...)$, where $F : [\mathbb{R} \times \mathbb{R} \times ...] \to \mathbb{R}$ is a measurable function. Assume that z_t has finite p^{th} moment for $p > 0$. Let $\boldsymbol{\acute{\epsilon}}$ denote a stochastic vector with the

same dimension and distribution as $\boldsymbol{\epsilon}_t$ such that $\acute{\boldsymbol{\epsilon}}$ and $\boldsymbol{\epsilon}_t$ are independent for all t. Then, for $k \geq 0$, the physical dependence $\delta_{kp}^{[z_t]}$ is defined as

$$\delta_{kp}^{[z_t]} = \left(\mathsf{E}\left(|F(\boldsymbol{\epsilon}_k, ..., \boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_{-1}...) - F(\boldsymbol{\epsilon}_k, ..., \acute{\boldsymbol{\epsilon}}, \boldsymbol{\epsilon}_{-1}...)|^p\right)\right)^{1/p}.$$

Step 2. Alternative Assumptions

In lieu of Assumption A9, one could impose more basic conditions on existence of finite p^{th} moments (for $p > 4$) for the innovations ψ_{qt} and ϕ_{it} (the primitive building blocks of our model) defined in Assumption A2. The requirements are as follows.

Assumption A10. *There exist constants p and A, with $p > 4$ and $0 < A < \infty$, such that*

$$\mathsf{E}\left(|\psi_{qt}|^p\right) \leq A, \quad \mathsf{E}\left(|\phi_{it}|^p\right) \leq A$$

for all $q = 1, ..., r$ and $i = 1, ..., N$.

We concurrently need a slight strengthening of Assumption A6 (on the characterisation of non-pervasiveness of the idiosyncratic component) to accommodate our upcoming proof. We mirror the relevant phrasing from Forni et al. (2017, Assumption 4) whereby restrictions are imposed on coefficients of the filter applied to the innovations driving the idiosyncratic component (rather than on the eigenvalues of the spectral density matrix of the idiosyncratic component).

Assumption A11. *There exist finite positive constants B, B_{is} for $i, s \in \mathbb{N}$, and $\rho \in [0, 1)$, such that*

$$\sum_{s=1}^{\infty} B_{is} \leq B \text{ for all } i \in \mathbb{N},$$
$$\sum_{i=1}^{\infty} B_{is} \leq B \text{ for all } s \in \mathbb{N},$$
$$|\beta_{is,k}| \leq B_{is}\rho^k \text{ for all } i, s \in \mathbb{N} \text{ and } k = 0, 1, ...$$

We make two remarks at this stage. First, we highlight that the time dependence of the idiosyncratic component is assumed to decline geometrically at common rate ρ. Second, we refer readers who may be interested in seeing how Assumption A11 automatically implies Assumption A6 to Forni et al. (2017, Proposition 1(i)).

To summarise, for the purpose of this appendix, we will replace Assumptions A9 and A6 from the main body of our text with newly-created Assumptions A10 and A11 respectively.

Step 3. Rate of Decline

We present the following result on the rate of decline of physical dependence under our model structure.

Proposition 2.10. *Under Assumptions A1–A5, A7–A8, and A10–A11, there exist constants* $\rho_1 \in (0,1)$ *and* $A_1 \in (0,\infty)$ *such that*

$$\mathsf{E}\left(|x_{it}|^p\right) \leq A_1, \quad \delta_{kp}^{[x_{it}]} \leq A_1 \rho_1^k$$

for all $i = 1, ..., N$.

Proof. Let us consider the first part of Proposition 2.10.

By the Minkowski inequality, we have

$$\left(\mathsf{E}\left(|x_{it}|^p\right)\right)^{1/p} = \left(\mathsf{E}\left(|\chi_{it} + \zeta_{it}|^p\right)\right)^{1/p} \leq \left(\mathsf{E}\left(|\chi_{it}|^p\right)\right)^{1/p} + \left(\mathsf{E}\left(|\zeta_{it}|^p\right)\right)^{1/p}.$$

Using the Minkowski inequality again (and absolute homogeneity of the norm),

$$\begin{aligned}
\left(\mathsf{E}\left(|\zeta_{it}|^p\right)\right)^{1/p} &= \left(\mathsf{E}\left(\left|\sum_{s=1}^{\infty}\sum_{k=0}^{\infty}\beta_{is,k}\phi_{s,t-k}\right|^p\right)\right)^{1/p} \\
&\leq \sum_{s=1}^{\infty}\sum_{k=0}^{\infty}\left(\mathsf{E}\left(|\beta_{is,k}\phi_{s,t-k}|^p\right)\right)^{1/p} \\
&\leq \sum_{s=1}^{\infty}\sum_{k=0}^{\infty}|\beta_{is,k}|\left(\mathsf{E}\left(|\phi_{s,t-k}|^p\right)\right)^{1/p} \\
&\leq A^{1/p}\sum_{s=1}^{\infty}\sum_{k=0}^{\infty}B_{is}\rho^k \\
&\leq A^{1/p}B\frac{1}{1-\rho},
\end{aligned}$$

where the penultimate and final lines follow from Assumptions A10-A11. An analagous inequality can be obtained for the common component χ_{it} using A2 and A10. Thus, the first part of Proposition 2.10 is established.

With respect to the second part of Proposition 2.10, we consider, for $k \geq 0$, the difference $\zeta_{ik} - \acute{\zeta}_{ik}$, where the constructed quantity $\acute{\zeta}_{ik}$ is defined as ζ_{ik} with ϕ_{s0} replaced with $\acute{\phi}_s$. Due to cancellations of all terms except at the point of discrepancy (at time 0), the difference simplifies to $\zeta_{ik} - \acute{\zeta}_{ik} = \sum_{s=1}^{\infty}\beta_{is,k}(\phi_{s0} - \acute{\phi}_s)$.

It follows from the Minkowski inequality and Assumptions A10 and A11 that

$$\begin{aligned}
\delta_{kp}^{[\zeta_{it}]} &= \left(\mathsf{E}\left(\left|\sum_{s=1}^{\infty}\beta_{is,k}(\phi_{s0} - \acute{\phi}_s)\right|^p\right)\right)^{1/p} \\
&\leq \sum_{s=1}^{\infty}\left(\mathsf{E}\left(\left|\beta_{is,k}(\phi_{s0} - \acute{\phi}_s)\right|^p\right)\right)^{1/p} \\
&\leq \sum_{s=1}^{\infty}|\beta_{is,k}|\left(\mathsf{E}\left(\left|(\phi_{s0} - \acute{\phi}_s)\right|^p\right)\right)^{1/p} \\
&\leq \rho^k\sum_{s=1}^{\infty}B_{is}\left(\mathsf{E}\left(\left|(\phi_{s0} - \acute{\phi}_s)\right|^p\right)\right)^{1/p} \\
&\leq \rho^k 2BA^{1/p}.
\end{aligned}$$

An analagous inequality can be obtained for the common component χ_{it} using A2 and A10. We then have, again by the Minkowski inequality,

$$\begin{aligned}\delta_{kp}^{[x_{it}]} &= \left(\mathsf{E}\left(|x_{it} - \acute{x}_{it}|^p\right)\right)^{1/p} = \left(\mathsf{E}\left(\left|(\chi_{it} - \acute{\chi}_{it}) + (\zeta_{it} - \acute{\zeta}_{it})\right|^p\right)\right)^{1/p} \\ &\leq \left(\mathsf{E}\left(|\chi_{it} - \acute{\chi}_{it}|^p\right)\right)^{1/p} + \mathsf{E}(|\zeta_{it} - \acute{\zeta}_{it}|^p))^{1/p} \\ &= \delta_{kp}^{[\chi_{it}]} + \delta_{kp}^{[\zeta_{it}]},\end{aligned}$$

from which the second part of the proposition follows.

This completes the proof of Proposition 2.10. □

Step 4. Establishing Assumption A9

We can use the C_r inequality for $r = 2$ to obtain

$$\begin{aligned}&\mathsf{E}\left[\max_{1\leq q\leq 2B_T+1}\left|\widetilde{g}_{\mathbf{x},is}^{B_T}(w_{j_q}) - g_{\mathbf{x},is}(w_{j_q})\right|^2\right] \\ &= \mathsf{E}\left[\max_{1\leq q\leq 2B_T+1}\left|\widetilde{g}_{\mathbf{x},is}^{B_T}(w_{j_q}) - \mathsf{E}\left(\widetilde{g}_{\mathbf{x},is}^{B_T}(w_{j_q})\right) + \mathsf{E}\left(\widetilde{g}_{\mathbf{x},is}^{B_T}(w_{j_q})\right) - g_{\mathbf{x},is}(w_{j_q})\right|^2\right] \\ &\leq 2\left[\mathsf{E}\left[\max_{1\leq q\leq 2B_T+1}\left|\widetilde{g}_{\mathbf{x},is}^{B_T}(w_{j_q}) - \mathsf{E}\left(\widetilde{g}_{\mathbf{x},is}^{B_T}(w_{j_q})\right)\right|^2\right] + \mathsf{E}\left[\max_{1\leq q\leq 2B_T+1}\left|\mathsf{E}\left(\widetilde{g}_{\mathbf{x},is}^{B_T}(w_{j_q})\right) - g_{\mathbf{x},is}(w_{j_q})\right|^2\right]\right],\end{aligned}$$

which is an expression in terms of the variance and the squared bias.

We can decompose the bias as

$$\begin{aligned}\mathsf{E}\left(\widetilde{g}_{\mathbf{x},is}^{B_T}(\omega)\right) - g_{\mathbf{x},is}(\omega) &= \sum_{h=-(T-1)}^{(T-1)}\left(1 - \frac{|h|}{T}\right)K\left(\frac{h}{B_T}\right)s_{\mathbf{x},is}(h)e^{-i2\pi\omega h} - \sum_{h=-\infty}^{\infty}s_{\mathbf{x},is}(h)e^{-i2\pi\omega h} \\ &\leq \left|\sum_{h=-(T-1)}^{T-1}\left(K\left(\frac{h}{B_T}\right) - 1\right)s_{\mathbf{x},is}(h)e^{-i2\pi\omega h}\right| \\ &\quad + \left|\sum_{h=-(T-1)}^{T-1}K\left(\frac{h}{B_T}\right)\frac{|h|}{T}s_{\mathbf{x},is}(h)e^{-i2\pi\omega h}\right| + \left|\sum_{|h|\geq T}s_{\mathbf{x},is}(h)e^{-i2\pi\omega h}\right|\end{aligned}$$

We can analyse the limit as $T \to \infty$ of each of the three terms and, as explained in Priestley (1982, p.459)) and Forni et al. (2017, Appendix A.), it can be shown

that the first term is $O(B_T^{-\kappa})$, the second is $O(T^{-1})$ and the third is $O(T^{-\kappa})$ so that the overall bias is dominated by the first term. Thereby, for the *squared* bias, a final rate of $O(B_T^{-2\kappa})$ is obtained.

The key part however, is the variance. We leverage Wu and Zaffaroni (2018, Theorem 4)[30] wherein it is proved that (under our given assumptions) there exists a constant, say m, such that

$$\mathsf{E}\left[\max_{1\leq q\leq 2B_T+1}\left|\widetilde{g}^{B_T}_{\mathbf{x},is}(w_{j_q})-\mathsf{E}\left(\widetilde{g}^{B_T}_{\mathbf{x},is}(w_{j_q})\right)\right|^2\right]\leq m\left(\frac{B_T\log B_T}{T}\right).$$

This completes the description of how Assumption A9 may be derived.

[30]The required result is obtained by setting $v=2$.

2.11 Appendix – Conditions for Convergence

The following discussion constitutes an addendum to Section 2.5.1 above.

By way of introduction, the EM algorithm is an ascent algorithm in the sense that the log-likelihood function is monotonically increasing. Ascent algorithms (Luenberger, 1984, Chapter 6; Lange, 2010, Chapter 15) can be analysed in a unified manner following a theory developed by Zangwill (1969). Wu (1983) showed that this general theory applies to the EM algorithm as defined in Section 2.5.1, as well as to some of its variants that Wu (1983) calls generalised EM.

In this appendix, we consider conditions for convergence of the EM algorithm based primarily on the seminal papers of Dempster et al. (1977) and Wu (1983). We also investigate the refinement of Balakrishnan et al. (2017), which harnesses so-called local strong concavity around the global maximum for its results.

2.11.1 Monotonicity of the EM algorithm

We begin with the celebrated monotonicity result of Dempster et al. (1977) which is that the likelihood function is non-decreasing in the number of EM iterations. Following the notation and derivations of Section 2.5.1, we have the expression whereby for $k = 0, 1, 2, ...,$

$$\begin{aligned}\ell\left(X;\boldsymbol{\theta}^{(k+1)}\right) - \ell\left(X;\boldsymbol{\theta}^{(k)}\right) &= \mathcal{Q}(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}) \\ &\quad - \left[\mathcal{H}(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}) - \mathcal{H}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\right],\end{aligned}$$

As noted previously, the first term is non-negative since $\boldsymbol{\theta}^{(k+1)}$ maximises $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ by construction. Further, the second term is non-negative since for any $\boldsymbol{\theta}$,

$$\begin{aligned}
\mathcal{H}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) - \mathcal{H}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) &= \mathsf{E}_{\boldsymbol{\theta}^{(k)}} \left[\log \frac{\mathcal{L}\left(F|X; \boldsymbol{\theta}\right)}{\mathcal{L}\left(F|X; \boldsymbol{\theta}^{(k)}\right)} \middle| X \right] \\
&\leq \log \mathsf{E}_{\boldsymbol{\theta}^{(k)}} \left[\frac{\mathcal{L}\left(F|X; \boldsymbol{\theta}\right)}{\mathcal{L}\left(F|X; \boldsymbol{\theta}^{(k)}\right)} \middle| X \right] \\
&= \log \int \frac{\mathcal{L}\left(F|X; \boldsymbol{\theta}\right)}{\mathcal{L}\left(F|X; \boldsymbol{\theta}^{(k)}\right)} \mathcal{L}\left(F|X; \boldsymbol{\theta}^{(k)}\right) dF \\
&= \log \int \mathcal{L}\left(F|X; \boldsymbol{\theta}\right) dF \\
&= \log 1 = 0,
\end{aligned}$$

which follows from Jensen's inequality and concavity of the logarithmic function. This was the subject of "Lemma 1" in Dempster et al. (1977). It follows that

$$\ell\left(X; \boldsymbol{\theta}^{(k+1)}\right) \geq \ell\left(X; \boldsymbol{\theta}^{(k)}\right),$$

for $k \geq 0$. The implication is that for a sequence of likelihood values $\{\mathcal{L}\left(X; \boldsymbol{\theta}^{(k)}\right)\}$ that is bounded from above[31], $\mathcal{L}\left(X; \boldsymbol{\theta}^{(k)}\right)$ converges monotonically to some L^* (Wu, 1983, p.96).

The monotonicity property of the EM algorithm guarantees that the procedure never worsens the likelihood of the data during its successive updates. This assures us that the algorithm does not diverge, although it does not answer the question of whether L^* is a global or local maximum, or even a stationary point.

2.11.2 Regularity Conditions

Given a d-dimensional vector of unknown parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ denotes the parameter space, and observations X, Wu (1983) assumes the following:

(i) $\boldsymbol{\Theta}$ is a subset in d-dimensional Euclidean space $\mathbb{R}^d$;

[31] A sequence $\{x_n\}$ is said to be bounded from above if there exists a positive constant M such that $|x_n| \leq M$ for all $n \in \mathbb{N}^+$.

(ii) $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathcal{L}(X;\boldsymbol{\theta}) \geq \mathcal{L}(X;\boldsymbol{\theta}_o)\}$ is compact for any $\mathcal{L}(X;\boldsymbol{\theta}_o) > -\infty$;

(iii) $\mathcal{L}(X;\boldsymbol{\theta})$ is continuous in $\boldsymbol{\Theta}$ and differentiable in the interior of $\boldsymbol{\Theta}$.

Above, the notation $\boldsymbol{\theta}_o$ represents an arbitrary value of $\boldsymbol{\theta}$.

As a consequence of conditions (i)–(iii), the sequence $\{\mathcal{L}\left(X;\boldsymbol{\theta}^{(k)}\right)\}$ for $k \geq 0$ is bounded from above for any $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}$, where in order to "avoid trivialities", the starting point $\boldsymbol{\theta}^{(0)}$ is assumed to satisfy $\mathcal{L}(X;\boldsymbol{\theta}^{(0)}) > -\infty$ (Wu, 1983, p.96–97).

Further, Wu (1983) assumes that $\boldsymbol{\theta}^{(k+1)}$ is a solution to the equation

$$\frac{\partial \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

for all $k \geq 0$. The interpretation is that each $\boldsymbol{\theta}^{(k)}$ is in the interior of $\boldsymbol{\Theta}$.

2.11.3 Theoretical Guarantees

Under the above regularity conditions, we have the following result.

Result Wu1. *Suppose $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\phi})$ is continuous in both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Then, all the limit points of any instance $\{\boldsymbol{\theta}^{(k)}\}$ of the EM algorithm are stationary points of $\mathcal{L}(X;\boldsymbol{\theta})$, and $\mathcal{L}(X;\boldsymbol{\theta}^{(k)})$ converges monotonically to some value $L^* = \mathcal{L}(X;\boldsymbol{\theta}^*)$ for some stationary point $\boldsymbol{\theta}^*$.* (Wu, 1983, "Theorem 2")

According to Wu (1983, p.98), the extra "continuity condition [that is required by Result Wu1] is very weak and should be satisfied in most practical situations. For convergence to stationary values it turns out to be the only required regularity condition (in addition to those given before). [Result Wu1] is most useful in that it covers a broad range of statistical applications."

While Result Wu1 is important, it does not rule out, for instance, convergence to a saddle point. To strengthen the guarantee, an additional restriction is needed. Indeed, under the same regularity conditions, along with the extra continuity condition mentioned in Result Wu1, we have the following result.

Result Wu2. *Suppose*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^*) > \mathcal{Q}(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*)$$

for any stationary point $\boldsymbol{\theta}^*$ *that is not a local maximiser of* $\mathcal{L}(X;\boldsymbol{\theta})$. *Then all limit points of any instance* $\{\boldsymbol{\theta}^{(k)}\}$ *of the EM algorithm are local maximisers of* $\mathcal{L}(X;\boldsymbol{\theta})$ *and* $\mathcal{L}(X;\boldsymbol{\theta}^{(k)})$ *converges monotonically to some value* $L^* = \mathcal{L}(X;\boldsymbol{\theta}^*)$ *for some local maximiser* $\boldsymbol{\theta}^*$. (Wu, 1983, "Theorem 3")

The interpretation of Result Wu2 is that if the $\mathcal{Q}$-function is not trapped at any point $\boldsymbol{\theta}_o$ that is a stationary point but not a local maximum of $\mathcal{L}(X;\boldsymbol{\theta})$ – which is the additional restriction required by Result Wu2 – then L^* is not only a stationary point but also a local maximum of $\mathcal{L}(X;\boldsymbol{\theta})$.

For the next available result, we take a slightly more informal approach. The reason is that the original theorems of Wu (1983) are phrased in respect of the generalised EM algorithm (GEM)[32], which we do not consider in our study. Indeed, a formal discussion of the GEM algorithm compels an introduction of terminology associated with so-called point-to-set maps (see, for instance, Douc et al. (2014, p.498) for a concise summary), and we believe that the added complexity is unnecessary for our current purpose. We simply note that the EM algorithm (the focus of our interest) is a special case of the GEM algorithm, and proceed with a heuristic exposition below.

The convergence of a sequence of likelihood values $\mathcal{L}(X;\boldsymbol{\theta}^{(k)})$ to some value L^* does not automatically imply the convergence of the corresponding sequence of iterates $\{\boldsymbol{\theta}^{(k)}\}$ to a point $\boldsymbol{\theta}^*$. The next result pertains to such iterates. Suppose all the conditions required by Result Wu2 hold.

Result Wu3. *If* $\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\| \to 0$ *as* $k \to \infty$ *and the set of local maxima*

[32]Depending on the application, a closed-form solution to the M step may not exist. If so, it may not be feasible to seek out the value of $\boldsymbol{\theta}$ that globally maximises $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$. Dempster et al. (1977) defined a procedure in which the M step only requires $\theta^{(k+1)}$ to be selected such that

$$\mathcal{Q}(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}).$$

In other words, $\boldsymbol{\theta}^{(k+1)}$ is chosen to increase the $\mathcal{Q}$-function above its value at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ rather than to maximise it over all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This procedure is the GEM algorithm.

with a given likelihood value is discrete, then $\boldsymbol{\theta}^{(k)}$ *converges to a local maximum.* (Wu, 1983, p.102)

A sufficient condition for $\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\| \to 0$ is that there exists a real positive constant, say $M_{\mathcal{Q}}$, such that

$$\mathcal{Q}(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) - \mathcal{Q}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) \geq M_{\mathcal{Q}} \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\|^2$$

for all k (Wu, 1983, "Condition 1"). Importantly, this condition is satisfied by the regular exponential family (Wu, 1983, p.101).

As regards the discreteness requirement, it is violated, for instance, if the likelihood has a ridge of stationary points in which it takes the value L^*.

Our final result of interest from Wu (1983) pertains to an important special case of a unimodal likelihood with a single stationary point. Assume the standard regularity conditions hold.

Result Wu4. *Suppose that* $\mathcal{L}(X; \boldsymbol{\theta})$ *is unimodal in* $\boldsymbol{\theta}$ *with* $\boldsymbol{\theta}^*$ *being the only stationary point in the interior of* $\boldsymbol{\Theta}$*, and that* $\frac{\partial \mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}$ *is continuous in both* $\boldsymbol{\theta}$ *and* $\boldsymbol{\phi}$*. Then, any EM sequence* $\{\boldsymbol{\theta}^{(k)}\}$ *converges to the unique maximiser* $\boldsymbol{\theta}^*$ *of* $\mathcal{L}(X; \boldsymbol{\theta})$. (Wu, 1983, "Corollary 1")

In other words, in this case, the EM procedure delivers the global maximum.

The selection of results from Wu (1983) presented above is enough for the purpose of our current work. Nevertheless, we discuss a few interesting extensions below.

2.11.4 An Interesting Extension

We refer the reader to McLachlan and Krishnan (2008) and Gupta and Chen (2011) for detailed historical overviews of the EM algorithm and its many extensions since the ground-breaking work of Dempster et al. (1977) and Wu (1983). The references therein are too numerous to cite here. We choose to focus instead on the relatively recent contribution of Balakrishnan et al. (2017), which we find interesting since it develops a rather general framework for characterising the

convergence of the EM algorithm.

To summarise our foregoing discussion, when the likelihood function is multimodal, the best guarantee provided by Wu (1983) is convergence to an arbitrary local maximum. A guarantee of this type does not preempt the EM algorithm from converging to a "poor" local maximum Balakrishnan et al. (2017, p.78), meaning one that is far away from any global maximum of the likelihood. Balakrishnan et al. (2017) addresses the concern above by outlining sufficient conditions for the EM algorithm to converge to a small neighbourhood of the global maximum. The paper explains:

> [...] our contribution is to provide a quantitative characterization of a basin of attraction around the population global optimum with the following property: if the algorithm is initialized within this basin, then it is guaranteed to converge to an EM fixed point that is within *statistical precision* of a global optimum. The statistical precision is a measure of the error in the maximum likelihood estimate other minimax optimal method; we define it more precisely in the sequel. Thus, in sharp contrast with the classical theory [...[33]] – which guarantees asymptotic convergence to an *arbitrary EM fixed point* – our theory guarantees geometric convergence to a *"good" EM fixed point.* (Balakrishnan et al., 2017, p.78, emphasis in original)

The result is valid for several canonical latent variable models and the conditions required pertain chiefly to strong concavity and first-order stability of the $\mathcal{Q}$-function. We discuss these next.

In respect of additional notation, suppose $\boldsymbol{\theta}^*$ is the true (unknown) parameter, let ∇ denote the gradient operator, and let $\langle \cdot, \cdot \rangle$ denote the inner product. Finally, define the Euclidean ball of radius $r > 0$ centred at $\boldsymbol{\theta}^*$ as

$$\mathbb{B}_2(r; \boldsymbol{\theta}^*) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq r\},$$

where $\|\cdot\|_2$ is the usual L_2-norm.

[33] References omitted.

The paper defines the following conditions to ensure that any sequence of iterates, when initialised in this ball, converges to a small neighbourhood of $\boldsymbol{\theta}^*$.

Condition Bala1. $\boldsymbol{\Theta}$ *is some non-empty convex set of parameters.*

Condition Bala1 ensures convexity of the domain.

Condition Bala2. *There exists some* $\lambda > 0$ *such that*

$$\mathcal{Q}(\boldsymbol{\theta}_1; \boldsymbol{\theta}^*) - \mathcal{Q}(\boldsymbol{\theta}_2; \boldsymbol{\theta}^*) - \langle \nabla \mathcal{Q}(\boldsymbol{\theta}_2; \boldsymbol{\theta}^*), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \leq -\frac{\lambda}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2$$

for all pairs $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{B}_2(r; \boldsymbol{\theta}^*)$

Condition Bala2 ensures λ-strong concavity *local* to $\boldsymbol{\theta}^*$ (i.e. over an r-neighbourhood of $\boldsymbol{\theta}^*$) . If one were to require, for instance, that Condition Bala2 held for all pairs $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, the condition would be referred to as *global* λ-strong concavity.

In order to present the main result of Balakrishnan et al. (2017) in a concise way, we introduce a new mapping operator at this stage to represent the EM update. We define $M : \boldsymbol{\Theta} \mapsto \boldsymbol{\Theta}$, such that

$$M(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\theta}'; \boldsymbol{\theta})$$

Using this notation, the EM algorithm given some initialisation $\boldsymbol{\theta}^{(0)}$, produces a sequence of iterates $\{\boldsymbol{\theta}^{(k)}\}$ for $k \geq 0$, where $\boldsymbol{\theta}^{(k+1)} = M(\boldsymbol{\theta}^{(k)})$.

A final condition, relating to first-order stability, is needed.

Condition Bala3. *There exists a sufficiently small* $\gamma \geq 0$ *such that*

$$\|\nabla \mathcal{Q}(M(\boldsymbol{\theta}); \boldsymbol{\theta}^*) - \nabla \mathcal{Q}(M(\boldsymbol{\theta}); \boldsymbol{\theta})\|_2 \leq \gamma \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2$$

for all $\boldsymbol{\theta} \in \mathbb{B}_2(r; \boldsymbol{\theta}^*)$.

Condition Bala3 is a first-order γ-stability condition which enforces the gradient maps $\nabla \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta})$ and $\nabla \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^*)$ to be close whenever $\boldsymbol{\theta}$ lies in an r-neighbourhood of $\boldsymbol{\theta}^*$. This will ensure that the mapped output $M(\boldsymbol{\theta})$ stays relatively close to $\boldsymbol{\theta}^*$.

We are now ready to present the key result of Balakrishnan et al. (2017).

Result Bala1. *Under Conditions Bala1–Bala3, for some radius $r > 0$ and $0 \leq \gamma < \lambda$, the map M is contractive over $\mathbb{B}_2(r; \boldsymbol{\theta}^*)$, in particular with*

$$\|M(\boldsymbol{\theta}) - \boldsymbol{\theta}^*\|_2 \leq \frac{\gamma}{\lambda} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2$$

for all $\boldsymbol{\theta} \in \mathbb{B}_2(r; \boldsymbol{\theta}^*)$; consequently, the EM iterates $\{\boldsymbol{\theta}^{(k)}\}$ converge to $\boldsymbol{\theta}^*$, namely

$$\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_2 \leq \left(\frac{\gamma}{\lambda}\right)^k \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\|_2$$

when initialisation $\boldsymbol{\theta}^{(0)} \in \mathbb{B}_2(r; \boldsymbol{\theta}^*)$. (Balakrishnan et al., 2017, "Theorem 4")

This completes our review (albeit limited in scope to results most relevant to us) of theoretical guarantees available for the EM algorithm.

"Is yesterday's model capable of explaining today's data?"
– Chu et al. (1996)

Chapter 3

Sequential Changepoint Detection via an Eigenvalue Ratio

We develop a procedure to detect changepoints in a static approximate factor model on a sequential basis. Specifically, we define a ratio of eigenvalues of the covariance matrix of N response variables. We compute this ratio each period using observations within a rolling window of size m over time and monitor the results. We declare a changepoint as soon as the value of this ratio breaches a pre-specified alarm threshold. To substantiate our procedure, we investigate the asymptotic behaviour (as $N, m \to \infty$) of our proposed ratio, and prove that, for specific eigenvalues, the ratio will spike upwards when a changepoint is encountered but not otherwise. Further, we propose a block-bootstrap procedure to obtain the alarm thresholds. Finally, we present simulation results and an empirical application based on FTSE 100 stock price data.

3.1 Introduction

3.1.1 Background

We characterise a changepoint as an instance in time in which a model undergoes a structural break. In other words, there arise abrupt differences in the values of model parameters before and after the changepoint. Our interest lies in the detection of changepoints in factor models for high-dimensional time series data.

Several authors have already recognised the importance of identifying changepoints of this nature. By way of example, we refer the reader to Breitung and Eickmeier (2011), Chen et al. (2014), Cheng et al. (2016), Corradi and Swanson (2014), Han and Inoue (2014), Yamamoto and Tanaka (2015), and Baltagi et al. (2017). These papers undoubtedly represent significant contributions in a growing field. However, they all focus on the setting where the focus is on the detection of structural breaks given a historical dataset. As such, the analyses are retrospective or *offline*.

In contrast, we aim to detect changepoints on a real-time basis. Thus, our approach is *online* in the sense that we examine the problem of sequential monitoring for a structural break when new data arrive steadily with each additional time period. As far as we are aware, with one important exception, namely Barigozzi and Trapani (2020), not many studies have considered the *online* case. It is in this area that our study hopes to make a positive contribution.

We believe that this is an important area of research since factor models are often used to exploit information in real-time. For instance, Doz et al. (2011) explains how factor models are used to forecast gross domestic product (GDP) for the recent past, present and near future using real-time data, a practice commonly referred to as "nowcasting". For such models to remain relevant over time, it is imperative that structural breaks brought about by fundamental policy, technological, economic, or news-related changes are detected accurately and in a timely fashion. This ensures that we can update model parameters when necessary to accurately reflect conditions in the prevailing market environment.

3.1.2 Summary

In this study, we explore the problem of timely detection of a change in the loadings of a static approximate factor model.

We explain (see Section 3.3) the well-known result that a model with a structural break in loadings for one of the factors is observationally equivalent to a model with stable loadings but in which one factor (or "pseudo-factor") emerges and another disappears. Effectively, we reduce the problem of detecting changes in loadings to one of analysing changes in the number of factors (or pseudo-factors) over time.[1]

We recall that the literature on estimation of the number of factors focusses on eigenvalues of the covariance matrix of the data (e.g. Bai and Ng (2002), Alessi et al. (2010), Ahn and Horenstein (2013), and Onatski (2010)). Even simple methods such as the celebrated scree-plot of Cattell (1966) are based on a visual analysis of eigenvalues. Given the discussion in the previous paragraph, it is perhaps not a surprise then that an eigenvalue-based strategy can also be used for changepoint detection (at least within our proposed framework). This is the motivation behind our approach. An advantage of our eigenvalue-based approach is that we can detect changepoints simply by estimating covariance matrices. We do not need to first estimate the full model (i.e. loadings, factors, etc.).

Our study is organised as follows. In Section 3.2, we provide a detailed survey of the vast time series literature on changepoint detection. In Section 3.3, we specify our model (with parameter instability) and outline a changepoint detection strategy based on an eigenvalue ratio. The strategy is summarised as follows: for a model containing r factors, where $1 \leq q \leq r$ factors undergo an abrupt change in loadings, we construct a ratio based on the $(r+1)^{th}$ eigenvalue of the population covariance matrix of N response variables. The ratio is defined such that the eigenvalue in the denominator is computed over a fixed window consisting of the first m initial observations. The eigenvalue in the numerator is initialised from

[1] It stands to reason that our techniques will also be immediately applicable to the problem of detecting changes whereby loadings remain stable but there is a sudden emergence of one (or more) new factor(s). We do not explicitly dwell on the latter case here.

data within a window covering the next m initial observations. In this regard, the first $2m$ observations are treated as a training sample. We propose a monitoring procedure whereby the numerator is then re-computed each period over a rolling window of m observations, whereas the denominator remains fixed at the beginning of the training sample. We propose the use of this eigenvalue ratio as a time-varying detection statistic, such that a changepoint is declared the first time this statistic breaches a pre-specified alarm threshold. In Section 3.4, we find (in Theorem 3.1) that, asymptotically, this eigenvalue ratio exhibits distinctive spiking behaviour upon encountering a changepoint but not otherwise. In Section 3.5, we propose an estimator for our eigenvalue ratio (and in Theorem 3.2, we establish the sample level counterpart to Theorem 3.1). In Section 3.6, we spell out the steps for implementing a sequential changepoint detection procedure in practice. In Section 3.7, we describe a bootstrap procedure in order to obtain critical values of our detection statistic for use as alarm thresholds. In Section 3.8, we present simulation results. In Section 3.9, we present an empirical application based on FTSE 100 data (Oct. 2013 to Oct. 2016). We are able to successfully detect a changepoint associated with the Brexit vote in the United Kingdom (UK) in June 2016. In Section 3.10, we reflect on window size. We outline possible directions for future study with respect to selection of m in an adaptive manner. We conclude in Section 3.11, and gather proofs in Section 3.12.

3.2 Survey of Changepoint Literature

Sequential changepoint detection refers to a class of methods for detecting changes in the state of a phenomenon with as low a delay as possible subject to a tolerable risk of a false alarm. The need for sequential monitoring for structural change in econometric analysis (and beyond) was pithily underscored in Chu et al. (1996, p.1045):

> In the real world, new data arrive steadily. Given a previously estimated model, the arrival of new data invites the question: is yesterday's model capable of explaining today's data?

Indeed, given the central importance of structural stability in the statistical analysis of time series, much effort has been devoted historically to research on sequential changepoint detection. The literature traces its roots right back to the manufacturing quality control techniques pioneered in the 1920s and 1930s (see Shewhart (1931)). These were followed by formal methods for sequential analysis such as the sequential probability ratio test (SPRT) of Wald (1948) and Wald and Wolfowitz (1948), and the cumulative sum (CUSUM) scheme of Page (1954), Lorden (1971), Moustakides (1986) and Ritov (1990). Detailed and technical textbook treatments containing many classical references for sequential changepoint detection are in Poor and Hadjiliadis (2008) and Tartakovsky et al. (2014), and key extensions to classical methods are reviewed in Polunchenko and Tartakovsky (2012) and Horváth and Rice (2014).

We already began our sketch of the academic landscape of this area in Section 1.2.2 above but our focus therein was on studies most closely related to our own (i.e. specifically in the context of factor models). In this section, we conduct both a broader and deeper review of the body of existing literature on sequential changepoint detection for time series. In particular, we briefly touch upon the classical origins of the field before proceeding to explore sequential detection in both the univariate or low-dimensional setting (Section 3.2.1) and the high-dimensional setting (Section 3.2.2). For the high-dimensional case, we also hone in on studies (mainly in the offline literature) that exploit the behaviour of the covariance matrix as does our own (Section 3.2.3). We end our review with a summary of the novel randomised sequential testing procedure of Barigozzi and Trapani (2020) and compare its approach with our own (Section 3.2.4).

Before embarking upon our survey, we remark that there is also a vast literature on Bayesian methods for sequential changepoint detection whose signature feature is that the changepoint is treated as a random variable possessing a prior distribution. This is in contrast to the non-Bayesian approach which posits that the changepoint is an unknown but not necessarily random quantity and even if it is random, its distribution is unknown. We briefly mention some key historical references in this paragraph but thereafter refrain from a discussion of Bayesian methods for changepoint detection. For despite their importance and historical

significance, they fall entirely outside the scope of our study. The Bayesian formulation was originally framed in Girshick and Rubin (1952) to solve a typical quality control problem for continuous technological processes. The first optimality results concerning Bayesian change detection algorithms were obtained in Shiryaev (1961, 1963, 1965). Further investigations can be found in Pollak and Siegmund (1985), Pollak (1985, 1987). We refer readers interested in additional details on developments in the Bayesian tradition to, for example, Basseville and Nikiforov (1993) and Fearnhead and Liu (2007).

3.2.1 Sequential Detection – Univariate & Small N Case

Classical Origins

Page (1954) proposed the following class of CUSUM schemes. Let Z_i be a function of the i^{th} sample statistic X_i and let $S_n = \sum_{i=1}^{n} Z_i, S_0 = 0$. A simple CUSUM scheme triggers a changepoint declaration at period

$$\inf \left\{ n : S_n - \min_{1 \leq j \leq n} (S_j) \geq c \right\},$$

where c is some critical threshold. In particular, when the chosen function is the log-likelihood ratio with two given densities, the CUSUM scheme is closely related to the repeated SPRT of Wald (1948) which harbours optimality properties (see Wald and Wolfowitz (1948)). For instance, say $X_1, ..., X_\kappa$ are iid with common density f_0 and $X_{\kappa+1}, X_{\kappa+2}, ...$ are iid with common density f_1 and if we let $Z_i = \log\left(f_1(X_i)/f_0(X_i)\right)$, then the Page (1954) scheme yields the stopping rule

$$\inf \left\{ n : \max_{1 \leq k \leq n} \left[\sum_{i=k}^{n} \log \left(\frac{f_1(X_i)}{f_0(X_i)} \right) \right] \geq c \right\}.$$

Stopping rules when parameters of the the post-change density, say $\theta \in \Theta$, are unknown and maximum likelihood estimates are used are typically referred to as

generalised likelihood ratio (GLR) schemes and bear the form

$$\inf\left\{n : \max_{1\leq k\leq n}\sup_{\theta\in\Theta}\left[\sum_{i=k}^{n}\log\left(\frac{f_\theta(X_i)}{f_0(X_i)}\right)\right]\geq c\right\}.$$

Most initial studies in statistical process control (such as those mentioned above) naturally focussed on independent data. Nevertheless, a large literature on monitoring procedures for univariate time series also developed. Some important works include Barnard (1959) which adapted the aforementioned CUSUM procedure for dependent data, and Basseville (1986) which studies sequential changepoint detection in autoregressive models and suggests alternative detectors based on Kullback divergence and Chernoff's distance between conditional laws. Lai (1995) not only contains an excellent summary of the state of the art of its time[2] but also goes on to develop a so-called "window-limited GLR" scheme. Specifically, Lai (1995) emphasises that there is a great computational complexity associated with the GLR scheme for non-independent observations and uses an idea put forth by Willsky and Jones (1976) carrying out only a limited number ($< n$) maximisations of the conditional log-likelihood given past observations to ameliorate computational burden. The Lai (1995) stopping rule (with size w window) is

$$\inf\left\{n > w : \max_{n-w+1\leq k\leq n}\sup_{\theta\in\Theta}\left[\sum_{i=k}^{n}\log\left(\frac{f_\theta(X_i|X_1,...,X_{i-1})}{f_0(X_i|X_1,...,X_{i-1})}\right)\right]\geq c\right\}.$$

[2]Also see Lai (2001).

Linear Models

We round up our summary above of the classical frontrunners in the field by introducing next the work of Chu et al. (1996). The latter is an influential study for several reasons, not least because it ushered research on sequential changepoint detection into the econometrics literature where the changepoint problem is often referred to as monitoring for structural breaks. Specifically, Chu et al. (1996) develops CUSUM schemes for so-called "recursive residuals" from a linear regression model[3] and exploits a functional central limit theorem with a Wiener process as a limit from which boundary crossing probabilities may be drawn.

Chu et al. (1996) is noteworthy in various additional respects. First, it heralded a paradigm shift; that is, classical procedures typically minimise the detection delay under some restrictions on the false alarm rate (e.g. average run length to false alarm). Such tests typically yield a false alarm with probability one. Instead, Chu et al. (1996) focusses on controlling asymptotic probability of Type I error if no change occurs and has asymptotic power one (i.e. guaranteed identification of a changepoint) under the alternative. In this sense, the studies in the tradition of Chu et al. (1996) discussed hereafter deviate from the classical approaches discussed in previous paragraphs.[4] Second, Chu et al. (1996) espoused the use of an initial training period of m observations in which it is assumed that no changepoint arises before monitoring purposefully begins (referred to in the paper as the "non-contamination" assumption). Our own procedure is grounded on the same idea – i.e. that we are effectively aiming to monitor stability of the historically adequate model. Third, Chu et al. (1996) focussed on the open-ended (infinite horizon) setting in which monitoring can proceed indefinitely if no changepoint is discovered. This is more suited to economics where marginal cost of sampling is negligible relative to, say, engineering. We too adopt the open-ended framework.[5]

[3] Chu et al. (1996) also proposes an alternative detector based on parameter fluctuations. Other studies along similar lines include Krämer et al. (1988) and Ploberger et al. (1989).

[4] Gösmann et al. (2022) contains a detailed explanation for the interested reader.

[5] Studies of monitoring procedures with a fixed endpoint include Leisch et al. (2000), Zeileis et al. (2005), Wied and Galeano (2005), and Dette and Gösmann (2020).

The literature on sequential detection for parameter changes in linear regression models was subsequently advanced by Horváth et al. (2004, 2007) which developed two classes of monitoring schemes to sequentially detect a structural change in a linear model after a training period of size m. The first class of procedures is based on weighted CUSUMs of residuals, in which the unknown in-control parameter has been replaced by its least-squares estimate from the training sample, whereas the second class of schemes makes use of the CUSUMs of recursive residuals. Kirch and Tadjuidje Kamgaing (2015) is an interesting study that explains how most of the detection statistics discussed heretofore can be expressed by means of estimating functions. This provides a unifying framework for the derivation of asymptotic results. Specifically, Kirch and Tadjuidje Kamgaing (2015) derives regularity conditions under which limit results for the standard CUSUM monitoring scheme as originally proposed by Chu et al. (1996) are obtained. The results unify several studies on linear models such as Horváth et al. (2004), Aue and Horváth (2004), and Hušková and Koubková (2005).

Further, Kirch (2008) investigates bootstrap methods to obtain critical values for sequential changepoint tests for changes in the mean of a univariate process with *iid* errors. Hlávka et al. (2012) develops a procedure for monitoring changes in the error distribution of autoregressive time series. Hlávka et al. (2016) combines the previous works by investigating bootstrap methods for sequential detection of changepoints in autoregressive models. Zhou et al. (2015) investigates sequential detection in the case of linear quantile regression models.

Non-Linear Models

Most initial studies following Chu et al. (1996) (such as alluded to above) remained focussed on linear regression models or linear dynamic models with weakly dependent data. Instead, Altissimo and Corradi (2003) considers the case of shifts in the mean of a possibly non-linear process, allowing for dependent and non-identically distributed data. Gut and Steinebach (2002, 2009) consider sequential detection in the case of renewal counting processes. Berkes et al. (2004) designs a detector for changes in the parameters of a generalised autoregressive conditional heteroscedasticity (GARCH) process. Berkes et al. (2004) relies on quasi-likelihood scores and does not use model residuals. Andreou and Ghysels

(2006) aims to monitor the stability of non-linear GARCH-type processes that exhibit volatility clustering by adapting the CUSUM scheme. Kulperger and Yu (2005) pursues an alternative approach based on partial sums of squared residuals of GARCH processes. Aue et al. (2006) considers a linear regression model with errors modelled by martingale difference sequences, which include heteroscedastic augmented GARCH processes. One method introduced is based on the CUSUM of the residuals and the other is based on squares of prediction errors.

Recent Developments

We mention a few select interesting avenues of recent research below.

Brodsky (2012) investigates the problem of sequential detection of changes in equations for unobserved state variables and observations of linear and nonlinear multivariate state-space models. Aue et al. (2012) develops techniques for use with high-frequency data. The authors take the view of a portfolio manager who has to decide in real-time whether to hold or sell the assets in her portfolio. Based on a training sample of size m, they define their procedures as first crossing times of a suitably constructed threshold function by a quadratic form detector built from least squares estimates of the portfolio betas. Fremdt (2015) argues that the performance of ordinary CUSUM procedures depends on the time of change and is best under early change scenarios. It thus proposes modified CUSUM procedures for the detection of abrupt changes in the regression parameters of multiple time series regression models that show a higher stability with respect to the time of change than ordinary CUSUM procedures. Gösmann et al. (2021) proposes a new sequential monitoring scheme for changes in the parameters of a multivariate time series. In contrast to procedures proposed in the literature which compare an estimator from the training sample with an estimator calculated from the remaining data, Gösmann et al. (2021) takes the novel approach of dividing the sample at each time point after the training sample. Estimators from the sample before and after all separation points are then continuously compared calculating a maximum of norms of their differences. Kirch and Stoehr (2022) derives the asymptotic distribution of detection delay times for sequential detection procedures based on U-statistics. It covers both a difference-of-means sequential test and a Wilcoxon sequential change point test which is found to

exhibit low detection delay in heavy-tailed settings.

Concluding Remarks

In view of natural limits to the scope of our present work relative to the sheer number of studies in this field, we concede that we cannot realistically aim to cover all aspects of the literature in this review.[6] Nevertheless, we have covered (to the best of our knowledge) the most salient features of the landscape and hope that we have sufficiently demonstrated the richness of the field as motivation for our upcoming work. We leave this section by remarking that sequential changepoint analysis has clearly seen an explosion of interest over the last decade and we expect that this will only increase in the near future due to the need, as noted in Horváth et al. (2021, p.2272), "to create changepoint detection tools suitable for complex, non-Euclidean data structures [such as networks] and high-dimensional and functional data". In the next section, we focus specifically on the high-dimensional case.

3.2.2 Sequential Detection – Large N Case

Early Studies

In contrast to the univariate and low-dimensional settings, the body of existing research thins when considering the high-dimensional case to the extent that one leading study, Chen et al. (2022, p.237), bemoans "a paucity of prior literature" in the area. Some of the first studies to tackle monitoring for a large number of data streams were Tartakovsky et al. (2006) and Mei (2010).

Tartakovsky et al. (2006) considers monitoring in the context of cyber-terrorism in the form of intrusion detection in multichannel sensor systems. A multichannel sensor system monitors N sequences of observations whose distributions follow a certain law up to some unknown instant; after this instant, one of the sequences changes its statistical properties. It is assumed that there exist data from N channels that are mutually independent but data within a channel are permitted

[6] A case in point is that the literature on the distinct but arguably related area of *offline* changepoint detection (which we have not mentioned) is even bigger than for the online case.

to be correlated or non-identically distributed. Tartakovsky et al. (2006, p.255) explains the motivation for a so-called "multi-chart" detection procedure as follows:

> [...] applying a multi-chart test, rather than a single-chart test, is vitally important. Otherwise, if a change occurs in one particular [channel] and all the data are mixed into one single statistic, this change will be obscured by the normal traffic in all other [channels]. This will result in a very large detection delay, and the attack may proceed undetected. On the other hand, if the attack is visible in several channels, [...] then the multi-chart tests proposed [...] are not optimal but they will still work.

In the case that pre- and post-change models are known, Tartakovsky et al. (2006, Equation 2) exploits the independence property across channels to easily construct a likelihood ratio based CUSUM statistic, say $U_i(t)$, where t denotes time and i denotes the channel. Then, the proposed detection statistic is

$$\inf \left\{ t \geq 1 : \max_{1 \leq i \leq N} U_i(t) \geq c \right\}$$

for some critical threshold c. In the (more realistic) case that the models are unknown, the parameters are estimated and the likelihood ratio statistic, $U_i(t)$, is replaced with one that is based on a score function using the logic that the latter should drift continually away from zero when encountering a change (see Tartakovsky et al. (2006, Equation 22)).

Mei (2010, p.419) expands upon the motivation for a so-called "global" detection procedure:

> Unfortunately, the local monitoring approach does not take advantage of global information, and may lead to large detection delays if several data streams provide information about the occurring event. More importantly, even if the local false alarm rate is well controlled at each data stream, the global false alarm rate can be severe when the number of data streams is large, leading to obvious costs and the

classic boy who cried wolf phenomenon.

Mei (2010) argues that the scheme of Tartakovsky et al. (2006) is best suited to the case where a single data stream is affected by a change, and is computationally infeasible when the number of affected data streams is even moderately large. Mei (2010) achieves a reduction in dimensionality by computing, say K, local CUSUM statistics based on all possible subsets of affected data streams that could arise. The proposed stopping rule is

$$\inf\left\{t \geq 1 : \sum_{k=1}^{K} V_k(t) \geq c\right\},$$

where $V_k(t)$ is a likelihood or score-based statistic like in Tartakovsky et al. (2006) but calculated at the local level. This rule is better-suited to detecting dense rather than sparse changes, but the assumptions (e.g. lack of cross-sectional dependence, and prior knowledge about possible subsets of affected streams) remain restrictive.

Xie and Siegmund (2013) assumes that data streams are independent and normally distributed with unit variance, and that the changes occur in their mean values. At each time period, a GLR statistic for each individual stream is computed under the assumption that a changepoint has occurred. The local GLR statistics are combined via a mixture model that has the effect of soft thresholding the local statistics according to a hypothesised fraction of affected data streams. The resulting local statistics are summed and compared with a detection threshold. A window-limited version of the procedure as in Lai (1995) is also developed. Chan (2017) is similar to Xie and Siegmund (2013) but with an improved choice of tuning parameters.

Recent Developments

A recent contribution to the state of the art is Chen et al. (2022)[7] which introduces a method for high-dimensional multiscale sequential changepoint detection in settings where an N-variate Gaussian data stream may undergo a change in mean (where each possible value of the mean in the post-change regime is referred to

[7]Also see the companion paper Chen et al. (2021, pre-print) which focusses on inference.

as a specific "scale"). The procedure works by performing likelihood ratio tests against simple alternatives of different scales for each individual data stream (which the paper refers to as a "coordinate"), and then aggregating test statistics across scales and coordinates. The method, titled "online changepoint detection (ocd)", is explained in Chen et al. (2022) roughly as follows:

Say there exists an N-dimensional independent Gaussian sequence $\mathbf{x}_t = \left(x_t^1, ..., x_t^N\right)'$ with each element distributed $\mathcal{N}(0, 1)$ under the pre-change regime. Chen et al. (2022) tests against a simple post-change alternative of $\mathcal{N}(b, 1)$ for some $b \neq 0$ in the j-th data stream. For $t \in \mathbb{N}, b \in \mathbb{R}\backslash\{0\}$ and $j \in \{1, ..., N\}$, the optimal procedure (following Page (1954)) is to declare a changepoint when the statistic

$$R_{t,b}^j = \max_{0 \leq h \leq t} \sum_{i=t-h+1}^{t} b(x_i^j - b/2)$$

exceeds a certain threshold. The logic is that $\sum_{i=t-h+1}^{t} b(x_i^j - b/2)$ may be viewed as a likelihood ratio test statistic between the null and the simple alternative using the tail sequence $x_{t-h+1}^j, ..., x_t^j$, and $R_{t,b}^j$ is the most extreme of these likelihood ratio statistics, over all possible starting points for the tail sequence. The paper refers to the $R_{t,b}^j$ for $j = 1, ..., N$ as "diagonal statistics" and recommends their individual use for detecting sparse changes (which have much of their signal concentrated in one coordinate).

As regards dense changes, the paper defines an ancillary statistic, say,

$$q_{t,b}^j = \arg\max_{0 \leq h \leq t} \sum_{i=t-h+1}^{t} b(x_i^j - b/2)$$

to denote the length of the tail sequence in which the associated likelihood ratio statistic (in the j-th coordinate) is maximised. Then, tail partial sums of length $q_{t,b}^j$ are computed in all other coordinates $j' \neq j$ given by

$$A_{t,b}^{j'j} = \sum_{i=t-q_{t,b}^j+1}^{t} x_i^{j'},$$

and these are aggregated to form what the paper refers to as an "off-diagonal" statistic anchored at coordinate j. These off-diagonal statistics are used to detect changes whose signal is not concentrated at a single coordinate. The diagonal and off-diagonal statistics are then aggregated by a proposed algorithm that is adaptive (to signal strength in a given coordinate) in nature.

The contributions of Chen et al. (2022) are sophisticated, efficient and perform well in numerical studies relative to Mei (2010), Xie and Siegmund (2013) and Chan (2017). Nevertheless, the theoretical framework, which stipulates mutually uncorrelated Gaussian observations that are independent over time, appears restrictive for high-dimensional time series applications often observed in macro- and financial econometrics. These limitations are acknowledged in the paper and conjecture is offered on how to extend the results (Chen et al., 2022, p.246) but no formalisation of the conjecture has been advanced as yet.

We now discuss some studies that allow for temporal and cross-sectional dependence. Groen et al. (2013) considers multiple series that co-break and argues that simultaneous examination of a set of series helps identify changes with higher probability or more rapidly than when series are examined on a case-by-case basis. It proposes a multivariate detector that takes the residuals from a set of equations recursively estimated over a monitoring period. The residuals are purged of cross-equation correlation and maximum and average CUSUM detection statistics are constructed. Groen et al. (2013) is significant because (unlike Tartakovsky et al. (2006) and Mei (2010), etc.), the focus is specifically on macroeconomic applications such as changes in the monetary policy regime or shifts in total factor productivity growth and so on, and cross-sectional dependence is permitted.

More recently, Manner et al. (2021) addresses challenges of high-dimensionality by using factor copula models which are able to capture dependence structure for a large number of data streams with a relatively low number of parameters. It compares rolling window parameter estimates to the parameter estimates of the factor copula over an initial training sample (over which no-instability is assumed). Estimation occurs via the method of simulated moments and the proposed detector is a moving sum (MOSUM) type statistic (see, for example,

Chu et al. (1995)) based on the quadratic form:

$$m^2 \left(\widehat{\theta}_{t-m+1:t} - \widehat{\theta}_{1:m}\right)' \left(\widehat{\theta}_{t-m+1:t} - \widehat{\theta}_{1:m}\right)$$

where m is a window size, $\widehat{\theta}_{1:m}$ denotes parameter estimates from a training sample of size m and $\widehat{\theta}_{t-m+1:t}$ represents parameter estimates from a rolling window of size m computed at time t. Manner et al. (2021) also provides a real-world financial risk management application of its procedure.

Gösmann et al. (2022) develops a sequential procedure capable of detecting changes in the mean vector of a successively observed high-dimensional time series with spatial and temporal dependence. The approach is based on a weighted CUSUM statistic for one-dimensional data (weighted roughly by the estimated long-run variance of an individual series), and the component-wise statistics are aggregated via checking for the maximum. Unlike Chen et al. (2022), in this paper, it is not the statistic itself that is greatly innovative since it follows in the footsteps of works such as Tartakovsky et al. (2006) and Mei (2010). However, this study is very interesting since under weak conditions on cross-sectional and temporal dependence, a suitably transformed version of the detection statistic is proved to converge in distribution to the range of a Brownian motion and this is, in turn, shown to be in the maximum domain of attraction of the Gumbel distribution. The latter is actually a result of independent interest in extreme value theory. The drawback, however, is that convergence rates in such limiting results are rather slow. As a consequence, Gösmann et al. (2022) also proposes a bootstrap procedure to approximate quantiles in practical applications.

Finally, we mention He et al. (2021), which investigates the problem of sequential changepoint detection in matrix-valued time series. The study uses the same intuition that we do (see Sections 3.2.4 and 3.3.2 below) – i.e. a factor model with instability can be expressed as a stable model with an expanded set of factors. Consequently, in the presence of a changepoint, the number of spiked eigenvalues in the second moment matrix of the data increases. He et al. (2021) proposes two families of procedures[8] - one based on the fluctuations of partial

[8]The randomisation scheme proposed differs to that of Barigozzi and Trapani (2020).

sums, and one based on extreme value theory - to monitor whether the first non-spiked eigenvalue diverges after a point in time in the monitoring horizon.

3.2.3 Changepoint Detection – Covariance Analysis

This particular section considers papers that analyse the covariance structure for changepoint detection in high-dimensions. There is a lack of studies (to our knowledge) that focus on *sequential* detection in high-dimensions by relying on the covariance matrix specifically. With the notable exception of Barigozzi and Trapani (2020), which we discuss separately in Section 3.2.4 below, and perhaps He et al. (2021) for matrix-valued series, studies in this area tend to target the offline case. For this reason, we include in our ongoing literature review select influential and/or recent works in this area for the *offline* case since the insights gained thereby are useful. We emphasise that this section does not intend to be fully comprehensive. Rather, the aim of this section is to provide a high-level overview of key available methodologies. For readers interested in further details, we can recommend the references contained within the papers cited in Section 1.2.2 and in those cited below.

One offline study (also previously mentioned in Section 1.2.2) that shares characteristics with our own is Baltagi et al. (2017). It exploits the exact same idea that we do (see Section 3.3.2 below for mathematical details); that is, a model with a break in loadings has an equivalent representation as an alternative model with stable loadings but an expanded set of factors (or pseudo-factors). Say there exist (historical) data on an N dimensional vector of time series observed over T periods. The data-generating process of Baltagi et al. (2017) is a static approximate factor model which undergoes a change in loadings immediately after period κ for $1 \leq \kappa < T$.

The suggested procedure of Baltagi et al. (2017) is as follows: (i) obtain a preliminary estimate, $\widehat{r}$, of the total number of factors ignoring any structural change; (ii) given $\widehat{r}$, use PCA (along with a suitable identification restriction) to estimate the pseudo-factors, say $\widehat{\mathbf{g}}_t$; (iii) split the data into a pre- and post-break

subsample for each candidate location of $\kappa \in \{1, ..., T-1\}$ as

$$\widehat{\Sigma}_{\mathbf{g},pre} = \frac{1}{\kappa} \sum_{t=1}^{\kappa} \widehat{\mathbf{g}}_t \widehat{\mathbf{g}}_t'$$
$$\widehat{\Sigma}_{\mathbf{g},post} = \frac{1}{T-\kappa} \sum_{t=\kappa+1}^{T} \widehat{\mathbf{g}}_t \widehat{\mathbf{g}}_t';$$

and finally, (iv) estimate the changepoint by least squares as

$$\widehat{\kappa} = \arg\min \left\{ \sum_{t=1}^{\kappa} \left(\text{vec} \left(\widehat{\mathbf{g}}_t \widehat{\mathbf{g}}_t' - \widehat{\Sigma}_{\mathbf{g},pre} \right) \right)' \left(\text{vec} \left(\widehat{\mathbf{g}}_t \widehat{\mathbf{g}}_t' - \widehat{\Sigma}_{\mathbf{g},pre} \right) \right) \right.$$
$$\left. + \sum_{t=\kappa+1}^{T} \left(\text{vec} \left(\widehat{\mathbf{g}}_t \widehat{\mathbf{g}}_t' - \widehat{\Sigma}_{\mathbf{g},post} \right) \right)' \left(\text{vec} \left(\widehat{\mathbf{g}}_t \widehat{\mathbf{g}}_t' - \widehat{\Sigma}_{\mathbf{g},post} \right) \right) \right\},$$

where $\text{vec}(\cdot)$ of an $m \times n$ matrix is the $mn \times 1$ vector obtained by stacking its columns. To summarise, as demonstrated at the beginning of Section 3.4 below, the instability leaves its signature in the covariance structure of the high-dimensional series and it is precisely this signature that both Baltagi et al. (2017) and we ourselves exploit.

Another offline study, Han and Inoue (2014, p.1117), is motivated by similar logic:

> Based on the fact that the presence of a structural change in factor loadings yields a structural change in second moments of factors obtained from the full sample principal component estimation, [...] our statistic compares the pre- and post-break subsample second moments of estimated factors.

Also see, for example, Corradi and Swanson (2014), which has similar foundations. In the studies cited above, it is the insightful use of the factor model structure (and the associated implications on the behaviour of leading eigenvalues of the covariance matrix) that facilitates a reduction in dimensionality.

Wang et al. (2021) is an interesting offline study insofar as it seeks to identify multiple changepoints in the covariance structure in a high-dimensional setting. It takes a different approach to papers we have previously referred to since it

does not rely on a factor model setup. The paper instead makes a generic assumption that we observe a sequence of independent and centred N-dimensional sub-Gaussian random vectors whose covariance matrices are piecewise constant. The dimensionality N is assumed to be larger than the sample size, say T, such that $N = O(T/\log T)$. A reduction in dimensionality is achieved by generating carefully chosen (see quotation below for details) univariate projections of the data using PCA. The paper then establishes optimality properties of applying the wild binary segmentation (WBS) algorithm (see Fryzlewicz (2014)) to the resultant one-dimensional series. The procedure is titled Wild Binary Segmentation through Independent Projections (WBSIP) and is explained in the authors' own words[9] as follows (Wang et al., 2017, p.13):

> The WBSIP algorithm is a generalization of the WBS procedure of [Fryzlewicz (2014)] for mean change point detection and further exploits the properties of shadow vectors. The WBSIP procedure begins by splitting the data into halves and by selecting at random a collection of M pairs of integers (s, e) [such that $1 \leq s < e \leq T$ and $e - s > N\log(T) + 1$]. In its second step, WBSIP computes, for each of the M random integer intervals previously generated, a shadow vector using one half of the data and its corresponding one-dimensional time series using the other half. The final step of the procedure is to apply the WBS algorithm over the resulting univariate time series.

Other offline studies that rely on binary segmentation (see Vostrikova (1981) and Venkataraman (1992)) are Aue et al. (2009) for a single changepoint and Cho and Fryzlewicz (2015) for multiple changepoints. In Aue et al. (2009), a CUSUM statistic is proposed for detecting and locating a single changepoint in the covariance structure of multivariate time series, and the extension to the detection of multiple changepoints via binary segmentation is only discussed heuristically.[10] Cho and Fryzlewicz (2015) goes further and develops a method for the multiple

[9] The quote is actually borrowed from a pre-print version of the paper since we were unable to obtain full access to the original article.

[10] The asymptotics in Aue et al. (2009) are for a fixed N and $T \to \infty$, so it does not specifically address the high-dimensional setting but we mention it in its capacity as a precursor to Cho and Fryzlewicz (2015).

changepoint case titled the "sparsified binary segmentation" procedure. The latter aggregates CUSUM statistics by adding only those that pass a certain threshold. It is precisely the thresholding step here that mitigates problems associated with high-dimensionality (Cho and Fryzlewicz (2015, p.476)):

> [...] only those temporal fragments of the CUSUMs that survive after the thresholding are aggregated to have any contribution in detecting and locating the change points. In this manner, we reduce the influence of those sequences that do not contain any change points so that the procedure is less affected by them, which can be particularly beneficial in a high dimensional context. Therefore, we can expect improved performance in comparison with methods without a similar dimension reduction step [...]

An innovative method is developed by Barigozzi et al. (2018a) which is also concerned with offline detection of multiple changepoints in the covariance structure of a high-dimensional time series. The paper combines a factor model structure (permitting both cross-sectional and temporal dependence) with earlier results from Cho and Fryzlewicz (2015) that establish that expectations of wavelet periodogram and cross-periodogram sequences of the common components have a one-to-one correspondence with the covariance structure of the input time series. In this way, any changepoint in the covariance structure of the common components is detectable as a changepoint in the means of the wavelet-transformed series. The procedure goes on to use a version of binary segmentation developed by Cho (2016) on the transformed series to identify changepoints.

Kao et al. (2018) develops a test for instability in the eigensystem of the covariance structure for a multivariate time series. The study leverages a strong invariance principle and a strong law of large numbers for the partial sample estimators of the covariance matrix, and uses these results to normalise a CUSUM-type test statistic, using Darling-Erdős limit theory (i.e. for the maximum of normalised sums of independent random variables).

Following Wang et al. (2021), Dette et al. (2022) also addresses the problem of offline estimation of a changepoint in the covariance matrix for a sequence of high-

dimensional vectors where the dimension is substantially larger than the sample size, but it focusses only a single changepoint. A two-stage approach is proposed to estimate its location. The first step consists of a reduction of the dimension to identify elements of the covariance matrices corresponding to significant changes, and the second step uses the components after dimension reduction to determine the position of the changepoint (via a CUSUM type statistic). In this study, dimension reduction is achieved neither via a factor model setup nor via PCA or thresholding, but through a non-standard procedure defined within the paper as part of its changepoint detection procedure.

We bring this overview to a close by mentioning the helpful review of Yu (2020) which summarises key technical results in offline changepoint detection. "Section 4.1" therein (Yu, 2020, p.19) addresses covariance analysis in the high-dimensional case with a particular focus on the methodology of Wang et al. (2021).

3.2.4 Sequential Detection – Barigozzi and Trapani (2020)

In the final part of our literature review, we focus on a single study closest to our own, Barigozzi and Trapani (2020). We highlight similarities relative to our work as well as salient points of divergence. For ease of reference, and for this section alone, we refer to Barigozzi and Trapani (2020) simply as "BT".

Shared Foundations

The model specification in BT is very similar to ours. That is, both studies postulate a static approximate factor model for an N-dimensional vector of (weakly stationary) time series, say $\mathbf{x}_t \in \mathbb{R}^N$, observed for up to T periods. The setting is one in which data points are continually gathered period-by-period and the data collection process may continue indefinitely (so $T \to \infty$). Both studies follow in the Chu et al. (1996) tradition, being concerned with sequential monitoring for an abrupt change in factor loadings that occurs at some point in time *after* an initial training period of size $m < T$. This is indeed the "non-contamination" restriction of Chu et al. (1996). Further, both studies define and make use of rolling covariance matrices with window size m.

We remark briefly that BT additionally considers a second type of changepoint which is the abrupt emergence of a new set of factors (see "$H_{A,2}$" therein). Such a change would certainly be detectable using our methodology too since one could argue, for instance, that the so-called emergent factors were always present but with loadings that changed from zero to non-zero at the changepoint. Thus, for the remainder of this section, we speak with reference to a break in the form of changes in loadings alone (or "$H_{A,1}$" in BT).

Moving forward, it is clear that both studies seek to exploit the same central premise that, say, an r-factor model with a break in loadings has an equivalent representation as an alternative model with stable loadings but an expanded set of factors. See, for instance, Sections 1.2.2 and 3.2.3 above, and Section 3.3.2 below. On this basis, both studies proceed as follows:

> Letting r be the number of common factors, we base our statistics on the fact that the $(r+1)$-th eigenvalue of the sample covariance matrix is bounded under the null of no change, whereas it becomes spiked under changes. (BT, p.5149)

The main modelling assumptions leading to the spiking behaviour alluded to above are only slightly different across the studies. BT, in its "Assumption 2", directly assumes that leading eigenvalues of the covariance of the common component diverge linearly with N and that all eigenvalues of the covariance of the idiosyncratic component remain bounded from above. Our study imposes more primitive assumptions in the style of Bai and Ng (2002) and Stock and Watson (2002a) but these lead to exactly the same result. In our study, this can be seen in the proofs of Lemma 3.4.1 and 3.4.2, while in BT, the approach we adopted is sketched out informally in the paragraphs immediately after "Assumption 2". These minor differences notwithstanding, both studies arrive at the same juncture; that is, our Lemma 3.4.3 corresponds with "Lemma 1" in BT.[11]

As regards estimation, both studies follow the approach of imposing a high-level

[11]For avoidance of doubt, we remark that there is a small difference in the convention used to represent the changepoint. In our study, the changepoint κ is defined as the final period in the pre-change regime. In BT, the changepoint τ is defined as the first period in the post-change regime. In other words, our κ equals $\tau - 1$ in the notation of BT.

condition that allows consistent recovery of the covariance structure of $\mathbf{x}_t$ using sample covariances. For the benefit of readers for whom more primitive conditions are of interest, both studies point to the fact that theorems linking the second moments of estimated covariances to second and fourth cumulants of the data are readily available in Hannan (1970, p.209–211). This approach is also shared by others in the literature (see, for example, Forni et al. (2009, p.1334)).

Given the framework and theoretical arguments described above, both studies propose online changepoint detection statistics based on the, say $(r+1)$-th, eigenvalue of the sample covariance matrix of the data. However, at this point, both studies are faced with the same challenge, which can be explained as follows:

> The main issue is that, under the null of no break, the $(r + 1)$-th sample eigenvalue does not have a known distribution, and indeed it cannot even be estimated consistently: [...] there is too much noise (due to N being large) to be able to identify the small signal coming from a bounded eigenvalue. (BT, p.5152)

Indeed, we agree that the distribution is unknown, and we find (see Lemma 3.5.1) that the estimation error for any eigenvalue is $O_p(N/\sqrt{m})$. BT finds a comparable rate (albeit not identical to ours since BT, following Trapani (2018), tackles the stronger problem of almost sure convergence).

This brings us to the point where the two studies take different approaches.

Key Differences

The chief distinction between our two studies may be summarised as follows. BT develops a formal online testing procedure for changepoint detection, whereas we develop what might be characterised a simple but useful diagnostic tool for online monitoring. We explain with some further details below.

BT sets up a statistic which has at its core the $(r+1)$-th estimated eigenvalue of the sample covariance matrix of the data. Clearly, such a statistic is of interest since (for reasons discussed above) it will be informative with respect to evaluating the null. Nevertheless, due to lack of results on consistency and on the limiting distribution under the null (also as discussed above), the statistic cannot

directly be used for hypothesis testing. Thus, BT recommends that the statistic of interest undergo a twice-randomisation procedure that is explained in detail in the paper (BT, p.5156–5160). The end result of the proposed procedure is that another statistic is generated, denoted in BT as Γ_t, which follows an asymptotic χ^2 distribution (with 1 degree of freedom) under the null of no change. Moreover, the statistic is independent across t conditional on the sample. Further, the information value of the original statistic with respect to the null is also preserved through the twice-randomisation procedure insofar as the abrupt change in the factor loadings translates into a shift in the mean of Γ_t. Thus, as BT explains (BT, p.5160), an online changepoint detector can be constructed based on the CUSUM of (a centred and standardised version of) of the sequence $\{\Gamma_t\}_{t=m}^{T}$.

The procedure of BT is innovative and sophisticated and represents a real breakthrough in the field. Nevertheless, the twice-randomisation procedure is relatively complicated and involves the choice of numerous tuning parameters (He et al., 2021, p.5). Our procedure is less sophisticated, by comparison, but we believe that it too is a useful contribution to the literature since it deftly exploits a simple intuition, which incidentally is also explained in BT as quoted immediately below. Note that the notation $\widehat{\lambda}^{(r+1)}(t)$ in the quotation below refers to the $(r+1)$-th eigenvalue of the rolling sample covariance matrix of the data at time t, while $\lambda^{(r+1)}(t)$ refers to the population counterpart. BT states:

> [...] it can be expected that $\widehat{\lambda}^{(r+1)}(t)$ may diverge to positive infinity even when $\lambda^{(r+1)}(t)$ is bounded; in this case, the divergence rate is $O(Nm^{-1/2})$, [...][12]. On the other hand, $\widehat{\lambda}^{(r+1)}(t)$ diverges at the faster rate $O(N)$ under the alternative. (BT, p.5156)

On the basis of the intuition above, we propose a suitably constructed eigenvalue ratio (see Section 3.5 below). We do not require consistency for all sample eigenvalues (i.e. at least not for eigenvalues that are not diverging linearly with N). In particular, we merely exploit the fact that the two orders of magnitude mentioned in the quotation from BT above are distinct. Indeed, all we need for a successful changepoint detection procedure is that our proposed ratio exhibits distinctive

[12] We excise here from the quotation any additional terms in the convergence rates found by BT that do not apply to our study.

(spiking) behaviour upon encountering a changepoint but not otherwise, and this is precisely what we establish in our study (see Theorem 3.2 below). Our argument is similar, for instance, to that made by Ahn and Horenstein (2013), which investigated ratios of adjacent eigenvalues of a sample covariance matrix under a factor model specification. Moreover, the argument is analogous to the one made by Lam and Yao (2012) albeit in a slightly different setting.[13] Further, since we do not have a limiting distribution available to us, we cannot directly find critical values under the null (e.g. by looking at theoretical quantiles of the distribution). Our answer to this problem is to propose a block-bootstrap procedure (suitable for time series data) to devise alarm thresholds.

We outline our proposed methodology in detail below.

3.3 Changepoint Model and Eigenvalue Ratio

3.3.1 Parameter Instability

We assume that all random variables belong to the Hilbert space $L_2(\Omega, \mathcal{F}, P)$ where $(\Omega, \mathcal{F}, P)$ is some given probability space. We begin with a real-valued double-indexed stochastic process of the form $\Xi = \{x_{it} |\ i \in \mathbb{N},\ t \in \mathbb{Z}\}$, where i denotes the cross-sectional index and t denotes the temporal index. Further, we let $\{\mathbf{x}_t = (x_{1t}, x_{2t}, ..., x_{Nt})' |\ t \in \mathbb{Z}\}$ be an N-dimensional subprocess of Ξ.

We impose a factor structure by assuming that $\mathbf{x}_t$ can be decomposed into a sum of unobserved common and idiosyncratic components.

$$\begin{aligned} \mathbf{x}_t &= \mathbf{c}_t + \mathbf{e}_t \\ &= \Lambda \mathbf{f}_t + \mathbf{e}_t \end{aligned} \tag{3.1}$$

[13] In the setting of Lam and Yao (2012), the factors are not only pervasive along the cross-sectional dimension but they also capture all serial correlation in the data. As such, identification is not based on the behaviour of the eigenvalues of the covariance matrix, but on the eigenvalues of the sum of autocovariance matrices. Nevertheless, the principles employed (i.e. in terms of exploiting a gap in the spectrum) are the same.

where

(i) $\mathbf{c}_t$ refers to the N-dimensional common component at time t;

(ii) $\mathbf{f}_t$ is an $r \times 1$ vector containing a single time series realisation for each of the $r \ll N$ common factors;

(iii) Λ is an $N \times r$ matrix of factor loadings; and

(iv) $\mathbf{e}_t$ refers to the N-dimensional idiosyncratic component at time t.

We extend this basic model by introducing parameter instability. Let us assume that there is a single abrupt break in the loadings matrix for $1 \leq q \leq r$ factors immediately after some common time period $\kappa \in \mathbb{Z}$, but right before the next time period $\kappa + 1$. We define the changepoint κ as the last time period under the pre-change regime. We impose this structural break as follows:

(i) let $(r-q)$-dimensional vector $\mathbf{f}_t^0$ represent the factors for which loadings, denoted by $N \times (r-q)$ matrix Λ^0, remain stable; and

(ii) let q-dimensional vector $\mathbf{f}_t^1$ represent the factors for which loadings change after κ; the pre-change $N \times q$ loadings matrix for these factors is denoted by Λ_a^1 and the corresponding post-change loadings matrix of identical dimensions by Λ_b^1;

(iii) let there exist two real-valued double-indexed stochastic processes of the form $\Xi^{(a)} = \{x_{it}^{(a)} |\ i \in \mathbb{N},\ t \in \mathbb{Z}\}$ and $\Xi^{(b)} = \{x_{it}^{(b)} |\ i \in \mathbb{N},\ t \in \mathbb{Z}\}$, and let N-dimensional subprocesses be denoted by $\{\mathbf{x}_t^{(a)} = (x_{1t}^{(a)}, x_{2t}^{(a)}, ..., x_{Nt}^{(a)})' |\ t \in \mathbb{Z}\}$ and $\{\mathbf{x}_t^{(b)} = (x_{1t}^{(b)}, x_{2t}^{(b)}, ..., x_{Nt}^{(b)})' |\ t \in \mathbb{Z}\}$ respectively.

With this additional notation, we can specify a model with a structural break:

$$\mathbf{x}_t = \begin{cases} \mathbf{x}_t^{(a)}, & t \leq \kappa \\ \mathbf{x}_t^{(b)}, & t > \kappa \end{cases} \tag{3.2}$$

where

$$\mathbf{x}_t^{(a)} = \Lambda^0 \mathbf{f}_t^0 + \Lambda_a^1 \mathbf{f}_t^1 + \mathbf{e}_t,$$

and

$$\mathbf{x}_t^{(b)} = \Lambda^0 \mathbf{f}_t^0 + \Lambda_b^1 \mathbf{f}_t^1 + \mathbf{e}_t.$$

For convenience, we also denote

$$\Lambda_a = \begin{bmatrix} \Lambda_0 & \Lambda_a^1 \end{bmatrix}; \Lambda_b = \begin{bmatrix} \Lambda_0 & \Lambda_b^1 \end{bmatrix}; \text{ and } \mathbf{f}_t = \begin{bmatrix} \mathbf{f}_t^{0\prime} & \mathbf{f}_t^{1\prime} \end{bmatrix}'.$$

Then, the common component $\mathbf{c}_t$ may be concisely expressed as

$$\mathbf{c}_t = \begin{cases} \Lambda_a \mathbf{f}_t, & t \leq \kappa \\ \Lambda_b \mathbf{f}_t, & t > \kappa \end{cases}.$$

We remark that model (3.1) is nested within model (3.2). It corresponds to the special case of no instability, in which the coefficients Λ_a and Λ_b are equal.

3.3.2 Alternative Representation

It is well-established (e.g. in Corradi and Swanson (2014), Han and Inoue (2014) or Baltagi et al. (2017)), that a model with a break in loadings has an equivalent representation as an alternative model with stable loadings but an expanded set of factors (or pseudo-factors).

In other words, model (3.2) has an alternative representation as follows:

For all $t \in \mathbb{Z}$,

$$\mathbf{x}_t = \Gamma \mathbf{g}_t + \mathbf{e}_t \tag{3.3}$$

where

$$\Gamma = \begin{bmatrix} \Lambda^0 & \Lambda_a^1 & \Lambda_b^1 \end{bmatrix}$$

with dimension $N \times (r+q)$;

$$\mathbf{g}_t = \begin{cases} A\mathbf{f}_t, & t \leq \kappa \\ B\mathbf{f}_t, & t > \kappa \end{cases}$$

with dimension $(r+q) \times 1$; and

$$A = \begin{bmatrix} I_{(r-q)} & 0_{(r-q)\times q} \\ 0_{q\times(r-q)} & I_q \\ 0_{q\times(r-q)} & 0_{q\times q} \end{bmatrix} \text{ with dimension } (r+q) \times r, \text{ and}$$

$$B = \begin{bmatrix} I_{(r-q)} & 0_{(r-q)\times q} \\ 0_{q\times(r-q)} & 0_{q\times q} \\ 0_{q\times(r-q)} & I_q \end{bmatrix} \text{ with dimension } (r+q) \times r.$$

I_{dim} and $0_{row\times col}$ are identity and zero matrices respectively where the subscripts indicate the relevant dimensions.

To summarise, instead of a model with r factors ($\mathbf{f}_t$) where q of the factors experience changes in loadings, we have an equivalent model with $r+q$ factors ($\mathbf{g}_t$) in which all factor loadings are in fact stable.

To see the equivalence between models (3.2) and (3.3), we simply observe that the common component is identical in both cases. That is, from model (3.2),

$$\begin{aligned} \mathbf{c}_t &= \begin{cases} \Lambda_a \mathbf{f}_t, & t \leq \kappa \\ \Lambda_b \mathbf{f}_t, & t > \kappa \end{cases} \\ &= \begin{cases} \Gamma A \mathbf{f}_t, & t \leq \kappa \\ \Gamma B \mathbf{f}_t, & t > \kappa \end{cases} \end{aligned}$$

so that $\mathbf{c}_t = \Gamma \mathbf{g}_t$ for all t, and model (3.3) follows.

The intuition is that in the alternative representation, our loadings instability is re-expressed in terms of two simultaneous offsetting changes in the number of factors. Specifically, q of the factors from the pre-change regime disappear and

are replaced in the post-change regime by a new set of q factors. Thus, in total, we see $r + q$ factors. Each of these is loaded uniquely and with coefficients that remain constant over time.

We note that a researcher may choose to refer to these emerging and disappearing factors ($\mathbf{g}_t$) as pseudo-factors since providing them with a reasonable interpretation may prove difficult. However, we believe that this is a decision best made depending on the context of any particular study. For brevity, in the remainder of this study, we no longer distinguish between the terms "factor" and "pseudo-factor" and refer to both $\mathbf{f}_t$ and $\mathbf{g}_t$ simply as factors.

3.3.3 Modelling Assumptions

Our model consists of two regimes, a pre-change regime when $t \leq \kappa$ and a post-change regime when $t > \kappa$. We make standard assumptions relating to static approximate factor models such as those considered in Bai and Ng (2002), Stock and Watson (2002a), and Forni et al. (2009) *within* each regime. Effectively, we assume that our model is piece-wise stationary.

In particular, we assume the following, for any $N \in \mathbb{N}$ and $t \in \mathbb{Z}$:

A *Common Component*

(i) $\mathsf{E}(\mathbf{f}_t) = 0_{N\times 1}$;

(ii) $\mathsf{E}(\mathbf{f}_t\mathbf{f}_t') = I_r$; and

(iii) $\mathbf{f}_t$ is second-order time-stationary.

(iv) As $N \to \infty$, the following conditions hold:

a. $(\Lambda_a'\Lambda_a/N) \to L_a$ where L_a is a positive definite $r \times r$ matrix;

b. $(\Lambda_b'\Lambda_b/N) \to L_b$ where L_b is a positive definite $r \times r$ matrix.

c. $(\Gamma'\Gamma/N) \to L_g$ where L_g is a positive definite $(r+q) \times (r+q)$ matrix.

B *Idiosyncratic Component*

(i) $\mathsf{E}(\mathbf{e}_t) = 0_{N\times 1}$;

(ii) $\mathsf{E}(\mathbf{e}_t \mathbf{e}_t{}') = \Sigma^{\mathbf{e}}$, where $\Sigma^{\mathbf{e}}$ is a positive definite matrix;

(iii) $\mathbf{e}_t$ is second-order time-stationary;

(iv) $\max_{i=1\ldots N} \sum_{j=1}^{N} |\mathsf{E}(e_{it} e_{jt})| \leq M < \infty$ for some positive constant M, where e_{it} and e_{jt} are individual elements of vector $\mathbf{e}_t$.

C *Orthogonality*

$$\mathsf{E}(\mathbf{f}_t \mathbf{e}_t{}') = 0_{r \times N}.$$

Collectively, assumptions A and B imply that the processes $X^{(a)}$ and $X^{(b)}$ are centred, non-degenerate and second-order time-stationary.

Moreover, Assumption A implies that non-zero eigenvalues of the covariance matrix of the common component diverge to infinity as $N \to \infty$ at a rate $O(N)$. This characteristic, sometimes referred to as the "pervasiveness" of factors, ensures that every factor has a non-negligible contribution to the variance of the response variables (see, for example, Bai and Ng (2002)).

Assumption B states that the maximum absolute column sum (hereafter referred to as the "column norm") of $\Sigma^{\mathbf{e}}$ is bounded from above for any N. This implies that all eigenvalues of $\Sigma^{\mathbf{e}}$ are bounded from above for any N. For our purposes, the latter is all that is needed and part (iv) of Assumption B may be relaxed. However, we use the proposed version of the assumption in accordance with the suggestion in (Bai and Ng, 2002, p.197). It is also the version relied upon by Fan et al. (2013). We maintain consistency with the literature.

3.3.4 Rolling Window

For convenience, and without loss of generality, we henceforth consider the point at which all our time series data begins as period $t = 1$. We now introduce the notation we will use to refer to a rolling window of observations on our N-dimensional vector $\mathbf{x}_t$. For some finite positive constant m such that $1 < m \leq t$,

we consider the matrix

$$\begin{bmatrix} x_{1t-m+1}, & \dots, & x_{1t} \\ x_{2t-m+1}, & \dots, & x_{2t} \\ \vdots & & \vdots \\ x_{Nt-m+1}, & \dots, & x_{Nt} \end{bmatrix}$$

as constituting an m-sized window of realised data under consideration at time t. When we fix window size m (predetermined by the researcher) and allow t to vary (from m to $m+1$ to $m+2$ and so on), we think of the above $N \times m$ panel as a window of observations that is rolling over time.

We always impose $m > 1$ because we require that there exist a minimum of two time series data points. Without data from a second time period, there would be no way to decide whether a changepoint exists in the first time period or not since there would be no basis for comparison. Further, we impose that monitoring may be expected to go on until period T for some $T > 2m$. We adopt a framework in which monitoring may continue indefinitely so that $T \to \infty$.

We define the $N \times N$ population rolling covariance matrix of the data as

$$\Sigma^{\mathbf{x}}_{[m]}(t) = \frac{1}{m} \sum_{h=t-m+1}^{t} \mathsf{E}(\mathbf{x}_h \mathbf{x}_h')$$

for $t = m, ..., T$, and we denote the s^{th} largest eigenvalue of this matrix by $\mu^{\mathbf{x}}_{s_{[m]}}(t)$ for $s = 1, ..., N$.

Similarly, we define a sample counterpart as

$$\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t) = \frac{1}{m} \sum_{h=t-m+1}^{t} \mathbf{x}_h \mathbf{x}_h'$$

for $t = m, ..., T$, and we denote the s^{th} largest eigenvalue of this matrix by $\widehat{\mu}^{\mathbf{x}}_{s_{[m]}}(t)$ for $s = 1, ..., N$.

We use analogous definitions for common components and idiosyncratic components. For instance, we define the $N \times N$ population rolling covariance matrix of the common and idiosyncratic components by $\Sigma^{\mathbf{c}}_{[m]}(t)$ and $\Sigma^{\mathbf{e}}_{[m]}(t)$ respectively.

Further, we denote the s^{th} largest eigenvalue of $\Sigma^{\mathbf{c}}_{[m]}(t)$ by $\mu^{\mathbf{c}}_{s_{[m]}}(t)$ and that of $\Sigma^{\mathbf{e}}_{[m]}(t)$ by $\mu^{\mathbf{e}}_{s_{[m]}}(t)$ for $s = 1, ..., N$.

We add another assumption as follows.

D *Changepoint Location*

> Changepoint κ, the final time period under the pre-change regime, is such that $\kappa \geq 2m$.

Assumption D ensures the availability of a reasonable sized set of pre-change data before sequential monitoring for changepoint detection purposefully begins.

Assumptions A-D are sufficient for a theoretical analysis of the asymptotic behaviour of a particular ratio of eigenvalues (to be defined shortly) when we let N tend to infinity. This will be the subject of Theorem 3.1.

3.3.5 Eigenvalue Ratio

We now arrive at the focal point of our study. That is, we propose the use of an eigenvalue ratio

$$\delta_{r+1_{[m]}}(t) = \frac{\mu^{\mathbf{x}}_{r+1_{[m]}}(t)}{\mu^{\mathbf{x}}_{r+1_{[m]}}(m)},$$

calculated each period t for $t = 2m, ..., T$. The denominator focusses on a fixed window consisting of the first m initial observations. Indeed, it is computed using the underlying matrix

$$\begin{bmatrix} x_{1,1}, & \dots, & x_{1,m} \\ x_{2,1}, & \dots, & x_{2,m} \\ \vdots & & \vdots \\ x_{N,1}, & \dots, & x_{N,m} \end{bmatrix}.$$

On the other hand, the numerator is re-computed each period on a rolling basis

by sequentially progressing through the matrices

$$\begin{bmatrix} x_{1,m+1}, & \dots, & x_{1,2m} \\ x_{2,m+1}, & \dots, & x_{2,2m} \\ \vdots & & \vdots \\ x_{N,m+1}, & \dots, & x_{N,2m} \end{bmatrix}, \begin{bmatrix} x_{1,m+2}, & \dots, & x_{1,2m+1} \\ x_{2,m+2}, & \dots, & x_{2,2m+1} \\ \vdots & & \vdots \\ x_{N,m+2}, & \dots, & x_{N,2m+1} \end{bmatrix}, ..., \begin{bmatrix} x_{1,T-m+1}, & \dots, & x_{1,T} \\ x_{2,T-m+1}, & \dots, & x_{2,T} \\ \vdots & & \vdots \\ x_{N,T-m+1}, & \dots, & x_{N,T} \end{bmatrix}$$

as the monitoring process unfolds. There is no overlap between observations considered in the numerator and denominator at any time period.

We can see now that Assumption D precludes the existence of a changepoint prior to period $2m$ to ensure two conditions: first, that the denominator is constructed from a valid comparator set of observations under the pre-change regime; and second, that we obtain a stable value for the numerator prior to monitoring in earnest for changepoints (in other words, allowing an opportunity for burn-in).

We show in Theorem 3.1 that as $N \to \infty$, the ratio $\delta_{r+j_{[m]}}(t)$ for $j = 1, ..., q$ tends to infinity at rate $\min\left\{\frac{t-\kappa}{m}, \frac{\kappa+m-t}{m}\right\} N$ so long as the rolling window covers at least one time period in both the pre- and post-change regimes; that is, so long as $t - m + 1 \leq \kappa$ and $t > \kappa$. If the rolling window contains only pre-change or post-change observations (but not both), the ratio remains bounded from above by a finite positive constant.

Theorem 3.1 serves as motivation for our changepoint detection methodology in the finite N case. We propose to monitor the behaviour of the $(r+1)^{th}$ eigenvalue using our eigenvalue ratio as a time-varying detection statistic. As soon as the value of this statistic breaches a pre-specified alarm threshold, we declare the preceding period a changepoint.

Thus, for a rolling window of fixed length m and with monitoring commencing at $t = 2m$, we have a changepoint estimator of the type

$$\inf\left\{t \geq 2m, \text{ such that } \delta_{r+1_{[m]}}(t) \geq H\right\} - 1$$

where H is the alarm threshold.

Asymptotic behaviour of the eigenvalue ratio is discussed in Section 3.4 below.

In practice, we need to make a few additional considerations in order to operationalise a sequential monitoring procedure. First, we require an estimator for our eigenvalue ratio. In Section 3.5, we propose an estimator based on eigenvalues of sample covariance matrices and provide a formal characterisation of its behaviour (in Theorem 3.2 therein). Second, we consider the question of what the appropriate length for the rolling window might be. Of course, there is also the question of how to estimate r. We provide brief informal guidance on these matters in Section 3.6 below.[14] Finally, we need to find alarm thresholds H. We propose the use of a bootstrap procedure based on the overlapping blocks resampling scheme of Kunsch (1989). We describe this in Section 3.7.

3.4 Behaviour of Eigenvalue Ratio

In this section, we establish the theoretical basis for the changepoint detection methodology we proposed in the previous section. As background, we note the well-known result (see, for instance, Trapani (2018) and Barigozzi and Trapani (2020)) that in a standard model with r factors and N response variables, the r largest eigenvalues of the covariance matrix of the response variables diverge to infinity as N goes to infinity (while the remaining eigenvalues remain bounded). We analyse how this result is affected in the presence of a changepoint.

We begin by considering the structure of $\Sigma^{\mathbf{x}}_{[m]}(t)$. Let us recall model (3.3):

$$\mathbf{x}_t = \Gamma \mathbf{g}_t + \mathbf{e}_t$$

where

$$\mathbf{g}_t = \begin{cases} A\mathbf{f}_t, & t \leq \kappa \\ B\mathbf{f}_t, & t > \kappa \end{cases} .$$

Under Assumptions B and C, $\mathsf{E}(\mathbf{x}_t \mathbf{x}_t{}') = \Gamma \mathsf{E}(\mathbf{g}_t \mathbf{g}_t{}')\Gamma' + \Sigma^{\mathbf{e}}$. Further, due to Assumption A, $\mathsf{E}(\mathbf{g}_t \mathbf{g}_t{}')$ equals AA' if $t \leq \kappa$, and BB' if $t > \kappa$.

[14] Also, see Section 3.10 for preliminary thoughts on developing an adaptive (data-driven) choice of window size.

Then,

$$\Sigma^{\mathbf{x}}_{[m]}(t) = \frac{1}{m}\sum_{h=t-m+1}^{t} \mathsf{E}(\mathbf{x}_h\mathbf{x}_h') = \Gamma\Sigma^{\mathbf{g}}_{[m]}(t)\Gamma' + \Sigma^{\mathbf{e}}_{[m]}(t) \tag{3.4}$$

where

$$\Sigma^{\mathbf{g}}_{[m]}(t) = \frac{1}{m}\sum_{h=t-m+1}^{t} \mathsf{E}(\mathbf{g}_h\mathbf{g}_h') \quad = \begin{cases} AA', & m \leq t \leq \kappa \\ \frac{m-(t-\kappa)}{m}AA' + \frac{t-\kappa}{m}BB', & \kappa < t \leq \kappa + m - 1 \\ BB', & \kappa + m - 1 < t \end{cases}.$$

In what follows, we denote two sets of generic finite positive constants by $\{\underline{M}_0, \underline{M}_1, ...\}$ and $\{\overline{M}_0, \overline{M}_1, ...\}$ where $\underline{M}_0 \leq \overline{M}_0, \underline{M}_1 \leq \overline{M}_1$, and so on. Specific values of the constants may change from line to line.

3.4.1 Eigenvalues of $\Sigma^{\mathbf{c}}_{[m]}(t)$

Lemma 3.4.1 characterises the behaviour of the $(r+j)^{th}$ eigenvalue, for $j = 1, ..., q$, of $\Sigma^{\mathbf{c}}_{[m]}(t) = \Gamma\Sigma^{\mathbf{g}}_{[m]}(t)\Gamma'$ in the following three cases: (i) the rolling window only contains time periods from the pre-change regime; (ii) the rolling window contains at least one time period from each of the pre- and post-change regimes; and (iii) the rolling window only contains time periods from the post-change regime.

Lemma 3.4.1. *Given model (3.3), under Assumptions A–D, there exists a finite positive constant* $\underline{M}_j$ *such that*

$$\mu^{\mathbf{c}}_{r+j_{[m]}}(t) \begin{cases} = 0, & m \leq t \leq \kappa \\ \geq \underline{M}_j \min\left\{\frac{t-\kappa}{m}, \frac{\kappa+m-t}{m}\right\} N, & \kappa < t \leq \kappa + m - 1 \\ = 0, & \kappa + m - 1 < t \leq T \end{cases}$$

for $j = 1, ..., q$.

Proof. (See Section 3.12.) □

3.4.2 Eigenvalues of $\Sigma^{\mathbf{e}}_{[m]}(t)$

Lemma 3.4.2 characterises the behaviour of the eigenvalues of $\Sigma^{\mathbf{e}}_{[m]}(t)$. We note that $\Sigma^{\mathbf{e}}_{[m]}(t) = \Sigma^{\mathbf{e}}$ for any $m \leq t \leq T$ so the behaviour of the eigenvalues remains the same across all three aforementioned cases of interest.

Lemma 3.4.2. *Given model (3.3), under Assumptions A–D, there exist finite positive constants* $\underline{M}_0$ *and* $\overline{M}_0$ *such that*

$$0 < \underline{M}_0 \leq \mu^{\mathbf{e}}_{N_{[m]}}(t) \leq ... \leq \mu^{\mathbf{e}}_{1_{[m]}}(t) \leq \overline{M}_0 < \infty.$$

Proof. (See Section 3.12.) □

Given Lemma 3.4.1 and Lemma 3.4.2, we can investigate the asymptotic behaviour of the $(r+j)^{th}$ eigenvalue, for $j = 1, ..., q$, of the covariance matrix of the response variables.

3.4.3 Eigenvalues of $\Sigma^{\mathbf{x}}_{[m]}(t)$

In Lemma 3.4.3 below, we find that $(r+j)^{th}$ eigenvalue, for $j = 1, ..., q$, diverges to infinity with N in the case in which the rolling window contains at least one time period from each of the pre- and post-change regimes. Instead, when the rolling window consists solely of time periods from either the pre- or the post-change regime (but not both), the $(r+j)^{th}$ eigenvalue, for $j = 1, ..., q$, remains bounded from above by a finite positive constant.

Lemma 3.4.3. *Given model (3.3), under Assumptions A–D, there exist finite positive constants* $\underline{M}_j$ *and* $\overline{M}_j$ *such that*

$$\mu^{\mathbf{x}}_{r+j_{[m]}}(t) \begin{cases} \leq \overline{M}_j, & m \leq t \leq \kappa \\ \geq \underline{M}_j \min\left\{\frac{t-\kappa}{m}, \frac{\kappa+m-t}{m}\right\} N, & \kappa < t \leq \kappa + m - 1 \\ \leq \overline{M}_j, & \kappa + m - 1 < t \leq T \end{cases}$$

for $j = 1, ..., q$.

Proof. (See Section 3.12.) □

3.4.4 Asymptotic Analysis of $\delta_{r+j_{[m]}}(t)$

We now evaluate the asymptotic behaviour of the eigenvalue ratio

$$\delta_{r+j_{[m]}}(t) = \frac{\mu^{\mathbf{x}}_{r+j_{[m]}}(t)}{\mu^{\mathbf{x}}_{r+j_{[m]}}(m)},$$

for any $j = 1, ..., q$. In Theorem 3.1, we establish that this ratio diverges to infinity with N when the rolling window contains at least one time period from each of the pre- and post-change regimes. Instead, when the rolling window consists solely of time periods from either the pre- or the post-change regime (but not both), the ratio remains bounded from above by a finite positive constant.

Theorem 3.1. *Given model (3.3), under Assumptions A–D, there exist finite positive constants* $\underline{M}_j$ *and* $\overline{M}_j$ *such that*

$$\delta_{r+j_{[m]}}(t) \begin{cases} \leq \overline{M}_j, & m \leq t \leq \kappa \\ \geq \underline{M}_j \min\left\{\frac{t-\kappa}{m}, \frac{\kappa+m-t}{m}\right\} N, & \kappa < t \leq \kappa + m - 1 \\ \leq \overline{M}_j, & \kappa + m - 1 < t \leq T \end{cases}$$

for $j = 1, ..., q$.

Proof. (See Section 3.12.) □

3.5 Estimation of Eigenvalue Ratio

The discussion in the previous section assumed that eigenvalues were known and the focus was on characterising the behaviour of the eigenvalue ratio as $N \to \infty$. In this section, we further consider estimation of the eigenvalue ratio and the

properties of our proposed estimator as both the dimensionality N and the sample size m go to infinity.

As regards estimation, we can compute the detection statistic

$$\widehat{\delta}_{r+1_{[m]}}(t) = \frac{\widehat{\mu}^{\mathbf{x}}_{r+1_{[m]}}(t)}{\widehat{\mu}^{\mathbf{x}}_{r+1_{[m]}}(m)},$$

where the numerator and denominator refer to eigenvalues obtained from sample rather than population covariance matrices.

Below, we use the notation $\sigma^{\mathbf{x}}_{ij_{[m]}}(t)$ and $\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t)$ to denote the generic $(i,j)^{th}$ (scalar) element of the population and sample covariance matrices respectively of N response variables constructed at time t from a time series of length m. Thereafter, we make the following additional assumption.

E *Covariance Matrix Estimates*

There exist positive finite constants ν_a and ν_b such that

(i) $m\mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t)\right)^2\right] < \nu_a$; and

(ii) $m\mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t)\right)^2\right] < \nu_b$.

Assumption E is not strictly necessary if we assume mild conditions on the auto-covariances and fourth cumulants of the response variables. We refer the reader to (Hannan, 1970, p.209-2011) for details. However, we follow the treatment in (Forni et al., 2009, p.1334), which imposes a requirement similar to Assumption E, since a derivation of the required conditions from first principles would not contribute much towards the specific aims of this study.

Under assumptions A–E, we establish in Lemma 3.5.1 below that if one scales an estimated eigenvalue $\widehat{\mu}^{\mathbf{x}}_{s_{[m]}}(t)$ by N then it is possible to obtain a bound for the estimation error that is vanishing at rate $m^{1/2}$. The result is valid for any combination of N and m and indeed for any $s = 1, ..., N$.

Lemma 3.5.1. *Given model (3.3), under Assumptions A–E, it holds that*

$$\left| \frac{\widehat{\mu}^{\mathbf{x}}_{s_{[m]}}(t)}{N} - \frac{\mu^{\mathbf{x}}_{s_{[m]}}(t)}{N} \right| = O_p\left(\frac{1}{\sqrt{m}}\right), \quad m \leq t \leq T,$$

for any $s = 1, ..., N$.

Proof. (See Section 3.12.) □

Lemma 3.5.1 effectively constitutes a consistency result as the sample size m goes to infinity for (scaled) eigenvalues that diverge linearly with N; and Lemma 3.4.3 indicates that the $(r+1)^{th}$ eigenvalue will diverge linearly with N provided m is increased in a specific way: (i) the rolling window (of size m) includes a sufficient number of time periods from the post-change regime for any spiking behaviour to manifest itself; and (ii) the rolling window (of size m) does not include so many time periods from the post-change regime that any pre-change signal is smothered out. In view of the above, and following Barigozzi and Trapani (2020, see Eq.14), we henceforth define $t^*_{N,m}$ as the point in time such that $t^*_{N,m} > \kappa$ and

$$\lim_{N,m\to\infty} \frac{N}{m}\left(t^*_{N,m} - \kappa\right) = \infty.$$

Similarly, we define $t^{**}_{N,m}$ as the period such that $t^*_{N,m} < t^{**}_{N,m} \leq \kappa + m - 1$ and

$$\lim_{N,m\to\infty} \frac{N}{m}\left(m - (t^{**}_{N,m} - \kappa)\right) = \infty.$$

The next result characterises how our detection statistic $\widehat{\delta}_{r+1_{[m]}}(t)$ exhibits distinct behaviour prior to a changepoint relative to once sufficient time has elapsed after it is encountered.

Theorem 3.2. *Given model (3.3), under Assumptions A–E, it holds that*

(i) $\widehat{\delta}_{r+j_{[m]}}(t) = O_p(1)$ *if* $m \leq t \leq \kappa$,

(ii) $\lim_{N,m\to\infty} \Pr\left(\widehat{\delta}_{r+j_{[m]}}(t) > \epsilon\right) = 1$ *for any* $\epsilon > 0$ *if* $\kappa < t^*_{N,m} \leq t \leq t^{**}_{N,m}$,

for $j = 1, ..., q$.

Proof. (See Section 3.12.) □

To gain intuition for the above result, it is useful to recall (once again[15]) the logic as summarised from Barigozzi and Trapani (2020, p.5156) that $\widehat{\mu}^{\mathbf{x}}_{(r+1)_{[m]}}(t)$ may diverge to positive infinity even when $\mu^{\mathbf{x}}_{(r+1)_{[m]}}(t)$ is bounded from above; in this case, the divergence rate is $O(N/\sqrt{m})$. On the other hand, $\widehat{\mu}^{\mathbf{x}}_{(r+1)_{[m]}}(t)$ diverges at the faster rate $O(N)$ when a changepoint is encountered. The same logic is also proffered by Lam and Yao (2012), albeit in a slightly different setting in which sums of autocovariance matrices (as opposed to covariance matrices) are the focus. Indeed, in the words of Lam and Yao (2012, p.704), "the estimation errors for the zero-eigenvalues is [*sic*] asymptotically of an order of magnitude smaller than those for the nonzero-eigenvalues." It is exactly this distinction in stochastic orders of magnitude that delivers us our result. (See also the theoretical justification for the eigenvalue ratio based method of Ahn and Horenstein (2013).)

3.6 Sequential Monitoring Procedure

In this section, we propose how to implement a method to detect changepoint κ based on the rolling window procedure analysed heretofore.

(i) First, we fix a value for the window length m. In practice, we find (based on simulations reported in Section 3.8 below) that, if adequate data are available, then choosing m such that $1 \leq m/N \leq 1.5$ works well.[16]

(ii) Given m, we define an initial training period given by the first $2m$ observations. Of course, we must be ready to make the assumption that there is no changepoint within the training period.

(iii) Given the training data, we obtain an estimate $\hat{r}$ of the number of factors using any standard method such as proposed in either Bai and Ng (2002),

[15] A direct quote from Barigozzi and Trapani (2020) was presented in Section 3.2.4 above.

[16] Proposed avenues of research on adaptive procedures which potentially allow data-driven choices of window size, m, are discussed in Section 3.10 below.

Alessi et al. (2010), Onatski (2010), or Ahn and Horenstein (2013).

(iv) We then sample data in a sequential manner until a changepoint is detected. For a rolling window of fixed length m and with monitoring commencing at $t = 2m$, we have a changepoint estimator of the type

$$\inf\left\{t \geq 2m, \text{ such that } \widehat{\delta}_{\hat{r}+1_{[m]}}(t) \geq H\right\} - 1$$

where H is a threshold whose computation we discuss in Section 3.7.

The rolling window methodology has the advantage that it is not computationally demanding. This is because we compute eigenvalues each period using only a fixed number of observations (no matter how long our sequential sampling continues).

We acknowledge that the methodology described above may be compromised by a poor initial estimate of r, in which case we would effectively run the risk of monitoring an irrelevant eigenvalue. However, we believe that this is a question for a separate branch of research (addressed by the papers on determination of the number of factors mentioned above). For our purposes, as a reasonable safeguard, we propose monitoring detection statistics corresponding to a set of eigenvalues around the $(\hat{r}+1)^{th}$ eigenvalue. (See, for instance, our approach in the empirical study in Section 3.9 below.)

3.7 Alarm Thresholds

In this section, we propose the use of a block bootstrap procedure in order to obtain critical values to use as alarm thresholds for our rolling window procedure.

3.7.1 Bootstrap Background

Subsequent to the phenomenal success of the bootstrap of Efron (1979) in providing answers to complex statistical questions involving independent data, it was soon established by Singh (1981) (see also Babu and Singh (1983)) that if one were to apply Efron's bootstrap to dependent data, inconsistency would ensue.

Indeed, the data-shuffling effect of resampling via Efron's bootstrap would mean that any information about the natural ordering of a sequence of observations and its underlying dependence structure would be entirely scrambled.

In general, block bootstrap methods are quite popular within time series contexts since by resampling blocks of consecutive observations, we are able to retain the dependence structure of the original data at least within any given block. Consistency can usually be restored through a judicious choice of block length – one that carefully balances the bias-variance tradeoff typically associated with estimators. While the idea of block-based sub-sampling methods can be traced back to Bartlett (1946), early breakthroughs in the specific area of bootstrapping were achieved by Hall (1985), Carlstein (1986), Kunsch (1989), and Liu and Singh (1992). See, for instance, Lahiri (2003) and Politis (2003), for excellent surveys of bootstrap techniques available for time series.

In our study, we choose to specifically deploy the overlapping blocks resampling scheme proposed in Kunsch (1989). Our choice is motivated by its simplicity, generality and successful use elsewhere in the literature on structural breaks within factor models. For instance, it is employed in Corradi and Swanson (2014) to obtain critical values for a test statistic proposed therein.

Finally, we remark that bootstrapping for factor models in time series is an open area of research (Gonçalves and Perron, 2014, 2020). While we pick a particular scheme in order to fix ideas, in principle, any procedure that works well would be suitable. In this regard, we briefly mention two other influential studies concerned with stationary time series which could potentially serve as alternatives to our choice. The first is the circular bootstrap of Politis and Romano (1992), which ensures that every observation from the original sample receives equal weight in being chosen for resampling.[17] The second is the stationary bootstrap of Politis and Romano (1994), which joins together blocks of random length – having a geometric distribution – and is able to generate bootstrap sample paths that are stationary series themselves. We do not consider the circular or stationary

[17]The goal is to mitigate the so-called "edge effects" faced by block-based schemes whereby observations at the start and the end of the sample are less likely to be resampled than those in the centre.

bootstraps in our study.

3.7.2 Threshold H

We describe the bootstrap of Kunsch (1989) for the case of our rolling window methodology. Our description follows the exposition in Lahiri (2003).

Let us consider the $N \times 2m$ panel of data that constitutes our training sample. Then, let $\mathcal{B}_\tau = (\mathbf{x}_\tau, ..., \mathbf{x}_{\tau+l-1})$ denote the block of length l starting with $\mathbf{x}_\tau$, for $1 \leq \tau \leq (2m - l + 1)$. The collection of all possible blocks of length l is $\{\mathcal{B}_1, ..., \mathcal{B}_{2m-l+1}\}$. We sample b of these blocks with replacement and refer to the resulting $(N \times bl)$ data matrix $(\mathcal{B}_1^*, ..., \mathcal{B}_b^*)$ as a bootstrap sample.

In practice, the bootstrap sample is typically chosen to be of the same size as the original sample. Thus, if b_0 denotes the smallest integer such that $b_0 l \geq 2m$, then we select $b = b_0$ blocks and use only the first $2m$ values to define the bootstrap sample. We thus obtain a trimmed bootstrap sample of dimension $N \times 2m$. As regards block length, we require that $l \to \infty$ and $l/m \to 0$ as $m \to \infty$. The former condition ensures that the bias of any bootstrap estimator is vanishing in the limit and the latter ensures that its variance is also vanishing as $m \to \infty$. Together, the two conditions imply convergence in mean-square, and by Chebyshev's inequality, we also have convergence in probability.[18] For our purposes, a simple function such as the square-root would satisfy the requisite conditions, but we can refer the interested reader to the study of Hall et al. (1995) for further guidance on optimal choice of block length in the block-based schemes of Carlstein (1986) and Kunsch (1989).[19]

We thereby present a methodology to obtain alarm threshold H. For a rolling window of fixed length m, block length of $l = \sqrt{2m}$, and monitoring now commencing at $t = 2m, 2m + 1, 2m + 2, ...$ and so on,

[18] This is the same bias-variance tradeoff that re-appears in many guises throughout the field of statistical inference. For example, we have already discussed it once at length in the context of lag-window estimation of the spectral density matrix in Chapter 2. The discussion therein related to choice of window size – see specifically Section 2.6.

[19] The guidance varies depending on what is the specific quantity we are trying to bootstrap.

(i) We generate B bootstrap samples for some large value of B using the overlapping blocks resampling scheme of Kunsch (1989) described above.

(ii) On each bootstrap sample, we calculate a bootstrapped version of our eigenvalue ratio denoted by $\widehat{\delta}^*_{\hat{r}+1_{[m]}}(2m)$. Of course, since critical values must be obtained under the null hypothesis of no instability, we must acknowledge the assumption that κ the changepoint – the final period under the pre-change regime – is such that $\kappa \geq 2m$.

(iii) Then, we choose H to be the $100(1-\alpha)^{th}$ percentile of the empirical distribution function of $\widehat{\delta}^*_{\hat{r}+1_{[m]}}(2m)$, where $\alpha \in (0,1)$ is some significance level.

Thus, we obtain a single threshold H from an initial period of training observations, and we use it for all time periods over which we monitor for a changepoint.

3.8 Simulations

We simulate data from the following model:

$$\mathbf{x}_t = \begin{cases} \mathbf{x}_t^{(a)}, & t \leq \kappa \\ \mathbf{x}_t^{(b)}, & t > \kappa \end{cases}$$

where

$$\mathbf{x}_t^{(a)} = \Lambda^0 \mathbf{f}_t^0 + \Lambda_a^1 \mathbf{f}_t^1 + \sqrt{\theta}\epsilon_t,$$

and

$$\mathbf{x}_t^{(b)} = \Lambda^0 \mathbf{f}_t^0 + \Lambda_b^1 \mathbf{f}_t^1 + \sqrt{\theta}\epsilon_t.$$

Λ^0 is an $N \times (r-q)$ matrix and Λ_a^1 and Λ_b^1 are both $N \times q$ matrices. $\mathbf{f}_t^0$ and $\mathbf{f}_t^1$ are $(r-q) \times 1$ and $q \times 1$ vectors respectively. All these items are constructed from draws of values from the $\mathcal{N}(0,1)$ distribution. ϵ_t is an $N \times 1$ vector drawn from the multivariate normal distribution $\mathcal{N}(0_{N\times 1}, \Sigma^\epsilon)$, where $\Sigma^\epsilon = (\beta^{|i-j|})_{ij}$ for $i,j \in [1,...,N]$ is a Toeplitz matrix with control parameter β. The constant θ allows us to control the proportion of the variance of $\mathbf{x}_t$ accounted for by the idiosyncratic component. Say we want this proportion to be π. Then we

set $\theta = [\pi/(1-\pi)][N^{-1}\sum_{i=1}^{N}\sigma^{\mathbf{c}}_{ii,t}]$ where $[N^{-1}\sum_{i=1}^{N}\sigma^{\mathbf{c}}_{ii,t}]$ is the average of the diagonal elements of the covariance matrix of the common component at time t. We assume all time series data begin at period $t = 1$.

3.8.1 Illustrative Example

We demonstrate a single run of our rolling window procedure using the example in Figure 3.1. For this particular example, we set model parameters:

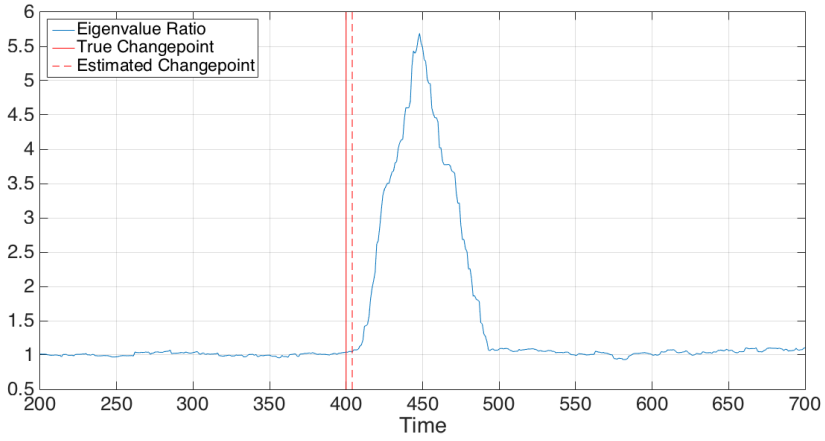
$$N = 100,\ \kappa = 400,\ r = 1,\ q = 1,\ \pi = 0.5,\ \beta = 0.25;$$

and control parameters for the detection procedure:

$$m = N,\ l = \lfloor\sqrt{2m}\rfloor,\ B = 500;$$

and then examine the path of the estimated eigenvalue ratio for the 2^{nd} eigenvalue.

Figure 3.1: $\kappa = 400$ and $\hat{\kappa} = 404$ (1% level)



We can consider a commonly-used measure to investigate performance of sequential changepoint detection procedures, the Average Run Length to False Alarm (ARL) (see, for example, Tartakovsky et al. (2014)). Given $\alpha = 0.05$, we expect the ARL to be $1/0.05 = 20$ periods. For $\alpha = 0.01$, we expect ARL to be $1/0.01 = 100$ periods.

We note that our changepoint is positioned at around 200 periods after monitoring commences. We thus expect to see around $200/20 = 10$ false alarms prior to the changepoint at the $\alpha = 0.05$ level and around $200/100 = 2$ false alarms at the $\alpha = 0.01$ level. In our example, we observe 12 and 2 false alarms respectively.

As regards the true changepoint, we detect it successfully with a delay of 4 periods (using the 1% alarm threshold). We find that the behaviour of our detection statistic is entirely in accordance with predictions from, say, Theorem 3.1 above.

Indeed, we observe distinctive upward spiking behaviour immediately upon encountering the changepoint. We also see that the spike becomes more pronounced as the rolling window slides forward and covers an increasing number observations from the post-change regime. The spike appears to reach its peak around the point where the rolling window straddles roughly an equal number of observations from both pre- and post-change regimes. Finally, as the rolling window continues to slide past the changepoint, thereby comprising an increasingly dominant share of post-change observations, the detection statistic reverts downwards. The ratio gradually returns towards its stable level (of around 1), to which it ultimately adheres once the pre-change signal is lost from the rolling window. All these findings appear to corroborate the presence of the $\min\left\{\frac{t-\kappa}{m}, \frac{\kappa+m-t}{m}\right\}$ factor in the overall divergence rate (as N) proffered by Theorem 3.1. In other words, the theory seems to be borne out numerically (at least for the given example).

To quantify detection delay, we compute a standardised measure by dividing through by the size of the rolling window. In this case, $m = 100$, so our standardised measure is $4/100 = 0.04$. The standardisation facilitates comparability across different window sizes. We move on now from our illustrative example in order to carry out a more extensive simulation analysis below.

3.8.2 Numerical Evidence on Detection Delay

In this section, we examine detection delays over simulations repeated 100 times.

With respect to control parameters, we let $N = 100$, $m \in \{50, 75, 100, 125, 150\}$, $r \in \{1, 2\}$, $q = 1$, $\pi \in \{0.25, 0.5, 0.75\}$, and $\beta \in \{0, 0.25\}$. Further, $l = \lfloor\sqrt{2m}\rfloor$

and $B = 500$. The changepoint is always set at $\kappa = 500$ and we monitor sequentially from period $t = 300$ until the changepoint is discovered or we hit period $T = 1000$.[20] Subsequently, we evaluate our estimators by looking at averages of the standardised detection delay performance statistic over all repetitions.

As in the individual illustration above, we have 200 periods prior to the changepoint, which means that with $\alpha = 0.01$, we expect on average 2 false alarms in each simulation. We apply a modification to the threshold (by dividing by 2 and then again by 100) in order to ensure that we can expect on average 1 false alarm over all 100 simulations. In other words, we impose a Bonferroni-style correction to oversee the family-wise risk of a false alarm.

Tables 3.1–3.3 provide breakdowns of frequencies (out of 100 simulations) of detection delays observed in low, medium and high noise settings under various scenarios. Specifically, Table 3.1 corresponds to the scenario in which a one factor model undergoes a structural break and there is no cross-sectional dependence in the idiosyncratic component. Table 3.2 corresponds to a one factor model that experiences a structural break in the presence of cross-sectional dependence in the idiosyncratic component. Table 3.3 corresponds to a two factor model that experiences a structural break for one of its factors in the presence of cross-sectional dependence in the idiosyncratic component. As previously noted, the cross-sectional dimension is fixed in all experiments to be $N = 100$ and results are reported for different window sizes, m, ranging from 50 to 150 in 25 period increments. In general, results appear to conform to expectations.

We summarise our findings below:

(i) Given the size of the cross-section, $N = 100$, the procedure exhibits somewhat erratic performance for relatively small window sizes, $m = 50$ and $m = 75$. This may be a manifestation of a poorly estimated covariance structure and/or insufficient data for the block bootstrap to be effective.

(ii) Performance of the procedure appears to be relatively stable for $m \geq 100$.

[20] Considering our maximum window size, $\overline{m}$, is 150, we stipulate that our monitoring commences at $t = 2\overline{m} = 300$. The choice of terminating at $T = 1000$ is arbitrary; we selected this specific value following Barigozzi and Trapani (2020, p.5166).

Indeed, not only do we observe a fall in the frequencies of premature termination of the procedure (i.e. due to false alarm) but we also observe broad improvements in the detection delay. We hence limit our focus in what follows to the cases which $m \geq 100$.

(iii) The procedure performs better under low and medium noise settings as compared with high noise settings. For example, in Table 3.1 for $m = 125$, the changepoint is detected within 10% of the size of the rolling window in 91–95% of experiments under the low and medium noise settings. The figure deteriorates to 75% in the high noise setting.

(iv) The procedure appears able to accommodate mild cross-sectional dependence (see Table 3.2) with only a minor deterioration in performance. Indeed, the figures in the previous bullet fall marginally to 87–94% under low and medium noise settings and 67% under the high noise setting.

(v) A final observation is that of a further deterioration in performance as we increase the number of factors in the data-generating process from 1 to 2. In other words, in Table 3.3, we now monitor the 3^{rd} (as opposed to the 2^{nd}) eigenvalue. The deterioration is particularly pronounced in the high noise setting. Interestingly, this finding is also discussed in Barigozzi and Trapani (2020, p.5168–5169), which explains that the power declines as r, the original number of factors, increases. Indeed, since the $(r+1)^{th}$ eigenvalue declines with r, one can expect signal to be increasingly dominated by noise as r rises.

Table 3.1: Detection Delay Frequency Tables
$\beta = 0, r = 1$

($\pi = 0.25$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	67	71	81	85	79
0.05–0.10	13	12	8	10	16
0.10–0.15	7	7	7	3	4
0.15–0.20	7	6	2	2	1
0.20–0.25	1	3	1	0	0
More than 0.25	1	0	0	0	0
Total	96	99	99	100	100

($\pi = 0.5$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	39	50	63	77	80
0.05–0.10	24	23	12	14	9
0.10–0.15	11	9	11	6	9
0.15–0.20	13	9	9	3	2
0.20–0.25	1	3	2	0	0
More than 0.25	7	1	1	0	0
Total	95	95	98	100	100

($\pi = 0.75$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	31	44	43	53	50
0.05–0.10	25	22	32	22	36
0.10–0.15	16	11	7	15	9
0.15–0.20	10	9	8	9	4
0.20–0.25	7	9	6	0	1
More than 0.25	8	4	3	0	0
Total	97	99	99	99	100

Table 3.2: Detection Delay Frequency Tables
$\beta = 0.25, r = 1$

($\pi = 0.25$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	51	57	67	76	71
0.05–0.10	18	14	26	18	19
0.10–0.15	12	11	2	4	6
0.15–0.20	8	7	1	1	3
0.20–0.25	8	5	1	0	0
More than 0.25	3	4	1	0	0
Total	97	98	98	99	99

($\pi = 0.5$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	32	47	56	70	68
0.05–0.10	30	29	25	17	21
0.10–0.15	13	11	12	9	5
0.15–0.20	11	10	2	1	3
0.20–0.25	9	1	2	1	2
More than 0.25	1	0	2	1	1
Total	96	98	99	99	100

($\pi = 0.75$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	29	35	37	42	45
0.05–0.10	19	32	36	25	19
0.10–0.15	19	12	25	19	13
0.15–0.20	15	8	1	11	11
0.20–0.25	8	6	0	1	9
More than 0.25	8	5	0	0	2
Total	98	98	99	98	99

Table 3.3: Detection Delay Frequency Tables
$\beta = 0.25, r = 2$

($\pi = 0.25$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	48	53	60	65	66
0.05–0.10	16	17	24	27	20
0.10–0.15	9	10	12	7	12
0.15–0.20	9	10	2	0	2
0.20–0.25	8	6	1	0	0
More than 0.25	8	3	1	0	0
Total	98	99	100	99	100

($\pi = 0.5$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	29	43	51	49	45
0.05–0.10	34	29	35	39	33
0.10–0.15	19	15	8	9	21
0.15–0.20	12	4	3	1	0
0.20–0.25	3	3	0	0	0
More than 0.25	1	4	1	0	0
Total	98	98	98	98	99

($\pi = 0.75$)

Standardised Delay	m=50	m=75	m=100	m=125	m=150
0.00–0.05	22	29	33	38	41
0.05–0.10	15	22	35	24	22
0.10–0.15	14	17	22	20	21
0.15–0.20	17	14	8	14	11
0.20–0.25	16	12	1	4	3
More than 0.25	13	4	0	0	0
Total	97	98	99	100	98

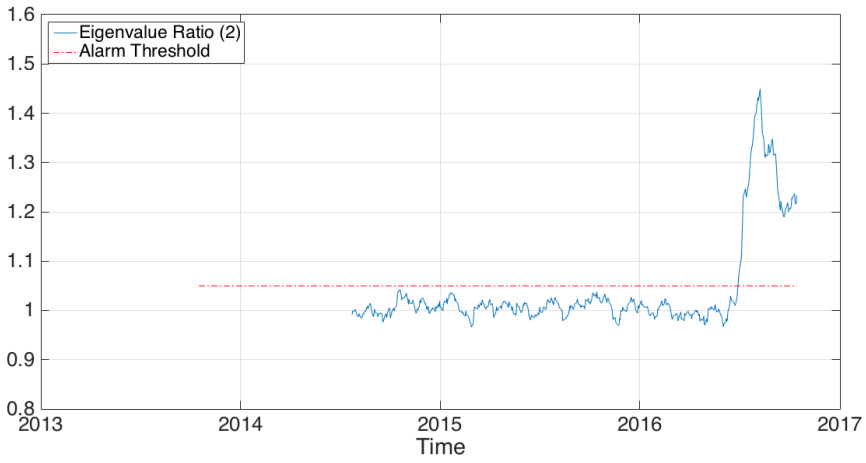
3.9 Empirical Application

We consider three years of daily stock price data[21] between 16 October 2013 and 15 October 2016 for FTSE 100 stocks. Of the 100 original series, we drop 9 series since they contained missing values for part of the period and thus cannot be used to estimate eigenvalues. We drop an additional 3 series because they contained erratic price movements very likely associated with data irregularities. We thus conduct our analysis based on the "adjusted closing price" for $N = 88$ stocks, from which we compute centred continuously-compounded returns.

We designate the first $z = 200$ time points as a training period and use the eigenvalue ratio method proposed by Ahn and Horenstein (2013) in order to estimate the number of factors.[22] We obtain an estimate of r equal to 1, and monitor eigenvalue ratios corresponding to the 2^{nd} eigenvalue onwards.

We deploy the bootstrap procedure for our rolling window methodology exactly as described in Section 3.7 above and obtain the results reported in Figures 3.2–3.4. The alarm thresholds are all associated with an $\alpha = 0.01$ significance level.

Figure 3.2: 2^{nd} eigenvalue: $\hat{\kappa} = 28/06/2016$



[21] The data were freely downloaded from "Yahoo! Finance" on 16 October 2016.

[22] We believe the method of Ahn and Horenstein (2013) is intuitive, simple-to-implement and closest in spirit to our own proposed procedure. In principle, any method may be used.

Figure 3.3: 3^{rd} eigenvalue: $\hat{\kappa} = 08/07/2016$

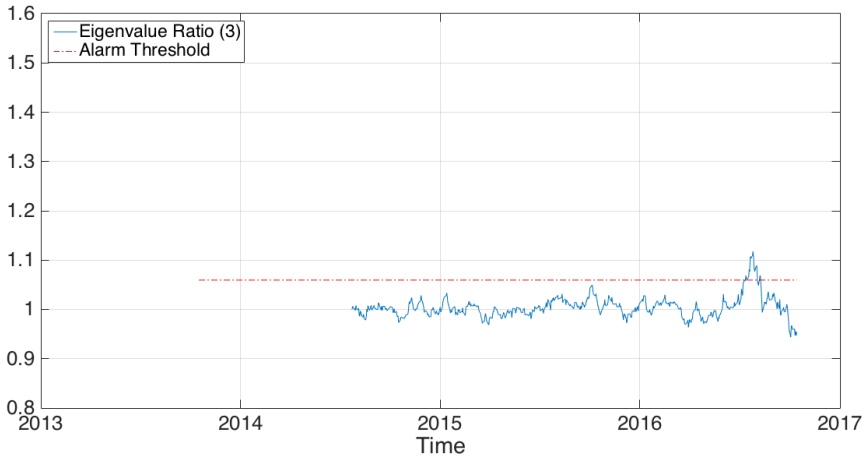
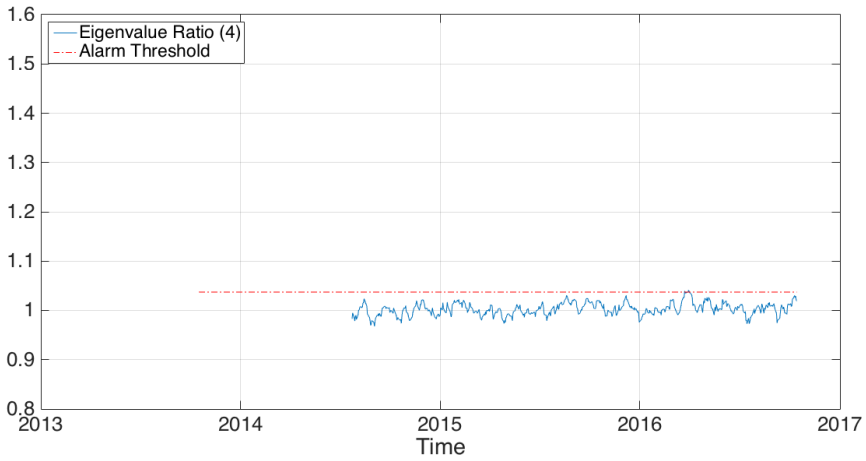


Figure 3.4: 4^{th} eigenvalue: $\hat{\kappa} = \varnothing$



We see from the graphs for the 2^{nd} and 3^{rd} eigenvalues that our detection statistic spikes and breaches the alarm threshold in the middle of 2016. In Figure 3.2, the estimated changepoint is 28/06/2016, and in Figure 3.3, it is 08/07/2016. We note that the signal for the instability contained in the 3^{rd} eigenvalue appears relatively weak in comparison to the 2^{nd} eigenvalue.

The period of turbulence identified above immediately follows the memorable date of 23/06/2016, on which the UK government, under Prime Minister David Cameron, held a referendum to determine whether or not to exit the European

Union. The following day, on 24/06/2016, markets reacted sharply to the news that the outcome of the referendum was one in favour of Brexit, an event that was widely regarded as a surprise. Indeed, Mr Cameron, who had staked his reputation on the "remain" campaign, resigned in the aftermath of the referendum. Our detection procedure appears to have successfully pin-pointed this period of stock market turbulence.

The eigenvalue ratio corresponding to the 4^{th} eigenvalue remains broadly stable over the entire period under consideration. While we do observe a breach of the alarm threshold in Figure 3.3, the dates of the breach, 22/03/2016–30/03/2016, as well as its small size, lack of persistence and isolated nature would suggest that this was a false alarm.[23]

Our results suggest it is possible that factor loadings may have changed on that day. It is also possible that one (or perhaps two) additional factor(s) had emerged. At any rate, if we were making use of factor models at the time, our detection procedure suggests that it would have been prudent to re-evaluate existing model parameters following Brexit. Notably, our detection delay appears to have been 5 days (or 5% of the size of the rolling window).

3.10 Reflecting on Choice of Window Size

Our rolling window approach can effectively be thought of as evaluating a discrepancy measure between two windows, one of which slides over the time series, say the "downstream" window, and another which remains fixed over an initial "non-contamination" period (Chu et al., 1996, p.1048), say the "upstream" window. At each time point, the discrepancy between the downstream and upstream windows with fixed width m is measured with the idea that peaks in the discrepancy measure may alert us about the existence of changepoints. See, for example, Gustafsson (2000), Kifer et al. (2004), Aminikhanghahi and Cook (2017), Bifet

[23]Interestingly, on 22/03/2016, Belgium was the target of three coordinated terrorist attacks by the Islamic State, claiming 32 lives and injuring more than 300 people. However, this is likely to be a coincidence as far as our analysis is concerned since there is no clear evidence of this event in the graphs for the other eigenvalues.

et al. (2017), and Truong et al. (2020) for discussion and surveys of methods based on sliding windows.

One concern with the standard sliding window approach as commonly employed – see, for instance, Xie and Siegmund (2013), Soh and Chandrasekaran (2017), Chan (2017), and Barigozzi and Trapani (2020) – is that it treats window size m as a parameter exogenously set by the investigator. Wide windows mostly take the global behaviour of observations into account, whereas narrow windows focus more on the local information. It is thus reasonable to expect that finite-sample performance of the procedure might depend on the precise choice of m. Rules-of-thumb for selecting m based on heuristics or numerical evaluations may be arbitrary and not always practical.

In this section, we put forward some proposals for making a data-driven choice of m in an adaptive manner. Implementation of these proposals and a thorough investigation of their theoretical properties is left as a matter for future research.[24]

Key Considerations

Let us consider the problem of choosing m in light of the insights from our foregoing theoretical analyses in Sections 3.4 and 3.5. On the one hand, we would like to choose a window size – our effective sample size – to be as large as possible. The benefit is that the larger is m, the more data are available for us to estimate the covariance structure. Further, the larger is m, the lower is the risk of false alarms (i.e. of detecting changepoints that do not exist). Conversely, the smaller is m, the less untenable is the non-contamination assumption for the upstream window. Moreover, the smaller is m, the greater is the sensitivity of our detection statistic to the presence of potential changepoints, and the lower is the expected delay in detection. A secondary benefit of a smaller m is that it accords an increase in computational speed with respect to calculating an $N \times N$ sample covariance matrix and its associated eigenvalues. A final consideration is that there may be an interest in maintaining recency, whereby older data are discarded either if they are deemed to be beyond a certain age threshold.

[24]The content of this specific section is best regarded as a work-in-progress. We nevertheless include it in this thesis merely to summarise the current state of our thinking.

Given the foregoing discussion, it is not *a priori* clear what is the appropriate choice of m. The only guidance to be gleaned from our asymptotic analysis on this matter is that $\sqrt{m}$ should grow no faster than N. So how can we proceed in practice? One way of introducing flexibility in choice of m is described below.

Adaptive Windowing

We first consider the adaptive window (ADWIN) of Bifet and Gavaldà (2007). Instead of choosing m, we choose $\underline{m}$ and $\overline{m}$, which denote minimum and maximum limits for the width of our window. We begin our monitoring procedure at time $t = 2\underline{m}$ with the statistic,

$$\widehat{\delta}_{r+1_{[\underline{m}]}}(t) = \frac{\widehat{\mu}^{\mathbf{x}}_{r+1_{[\underline{m}]}}(t)}{\widehat{\mu}^{\mathbf{x}}_{r+1_{[\underline{m}]}}(\underline{m})},$$

which is based on a small window. Then, for each additional pair of data points received, we increase our window size by one unit. For instance, at time $t = 2\underline{m} + 1$, the upstream window would remain fixed over the $\underline{m}$ points $\{1, ..., \underline{m}\}$ and the downstream window would slide one period to cover the m points $\{\underline{m} + 2, ..., 2\underline{m} + 1\}$. At time $t = 2\underline{m} + 2$, both windows would simultaneously expand. Indeed, the upstream window would now contain the $\underline{m} + 1$ points $\{1, ..., \underline{m} + 1\}$ and the downstream window would contain the $\underline{m}+1$ points $\{\underline{m}+2, ..., 2\underline{m}+2\}$.[25] In other words, we commence sequential monitoring with windows of minimum size but progressively extend their width until either the maximum size limit, $\overline{m}$, is reached or a change is detected. If the maximum size is reached, we discard the historically old data and recommence monitoring based only on the most recent set of $2\underline{m}$ observations available.

By way of justification, we quote from Bifet and Gavaldà (2007):

> The window will grow automatically when the data is stationary, for greater accuracy, and will shrink automatically when change is taking place, to discard stale data. This delivers the user or programmer from having to guess a time-scale for change.

[25] By comparison, at time $t = 2\underline{m} + 2$ in the original scheme, the upstream window would cover the set $\{1, ..., \underline{m}\}$ and the downstream window the set $\{\underline{m} + 3, ..., 2\underline{m} + 2\}$.

> Contrary to many related works, we provide rigorous guarantees of performance, as bounds on the rates of false positives and false negatives. In fact, for some change structures, we can formally show that the algorithm automatically adjusts the window to a statistically optimal length. (Bifet and Gavaldà, 2007, p.443)

While the ADWIN procedure is adaptive, it still involves the choice of minimum and maximum tuning parameters, and thus retains some of the arbitrariness of the fixed width approach. We discuss an alternative strategy below.

Information Criterion for Window Size Selection

Another possibility is to consider the choice of m in light of insights from papers that have employed model selection techniques in factor model settings. See, for example, Bai and Ng (2002), Hallin and Liška (2007), and Alessi et al. (2010), which propose various information criteria – i.e. in a bid to balance the tradeoff between goodness-of-fit and parsimony – for determining the number of common factors. We draw inspiration from this literature to ask whether we too can construct an information criterion which takes into account, for our purposes, the potential tradeoff between quality of estimation of the covariance structure and speed of detection of the changepoint.

For instance, we could exploit the notion that for any given eigenvalue, our ratio should (at least in the population) be close to unity when there is no instability. To fix ideas, suppose we are prepared to assume that the first S observations satisfy the non-contamination assumption of Chu et al. (1996), and we wish to choose from a set of candidate window sizes, $\{\underline{m}, ..., \overline{m}\}$, where $1 \leq \underline{m} < \overline{m} \leq \frac{1}{2}S$. It may then be possible to use such a training sample to inform our choice of m by assessing which of the candidate values is able to minimise a penalised mean-square error criterion, say, $IC_N(m)$. Consider, for instance, the formulation

$$\begin{aligned} m^* &= \underset{\underline{m} \leq m \leq \overline{m}}{\arg\min}\, IC_N(m) \\ &= \underset{\underline{m} \leq m \leq \overline{m}}{\arg\min} \left\{ \mathcal{V}(m) + m\mathcal{P}(N) \right\}, \end{aligned}$$

where we can expect the mean-square error,

$$\mathcal{V}(m) = \frac{1}{S - 2\overline{m}} \sum_{t=(2\overline{m}+1)}^{S} \left(\widehat{\delta}_{(r+1)_{[m]}}(t) - 1 \right)^2,$$

to be minimised for $m = \overline{m}$ (the maximum size allowed), and hence, we introduce the penalty function, $\mathcal{P}(N)$, so as to avoid over-sizing the window. In respect of penalisation, we could perhaps consider use of the Akaike Information Criterion (AIC) and/or the Bayesian Information Criterion (BIC), with penalties taking the form of multiples of $2/N$ and $(\log N)/N$ respectively (Bai and Ng, 2002, p.202). Of course, a thorough investigation of the feasibility of this approach and how it may affect the asymptotic properties of our detection statistic is warranted.

Other Possibilities

A further possibility may be to simultaneously deploy an ensemble of windows of different sizes. The advantage would be that we could combine information from multiple detectors, each with a different sensitivity to local behaviour, so that they all serve to mutually corroborate each other with the goal of balancing speed of detection against risk of false alarms. Such a strategy is inspired by the multi-scale procedure found in Chen et al. (2022) which concurrently monitors for so-called sparse versus dense changepoints and combines results from multiple detectors to generate an adaptive procedure.

Another option is to look into cross-validation techniques, although their suitability for time-ordered dependent data is an open area of research (Bergmeir and Benítez, 2012). Recent advances have been made, for instance, by Bergmeir et al. (2018), which shows that for purely autoregressive models, the use of standard K-fold cross validation is possible provided the models considered have uncorrelated errors.[26]

Finally, we mention the historically influential study of Hall et al. (1995), which minimises a mean-square error criterion to determine optimal block length in

[26] We are also aware of the work-in-progress of Pellegrino (2022), which investigates a generalisation of the delete-d-jackknife whereby any data points removed are suitably interpolated so that the overall time series structure is preserved.

the case of the Carlstein (1986) and Kunsch (1989) bootstrap schemes. The study finds that the optimal asymptotic formula for block length, say l, is $l = O(m^{1/k})$, where m is the number of available time series observations, and $k = 3, 4$, or 5 depending on context (one of which pertains to estimation of quantiles of a distribution). The paper validates its theoretical results numerically for autoregressive and moving average data-generating processes. It would be worth exploring if any of the insights therein could be extended for use in our case.

We re-emphasise that the nascent ideas in this section remain largely unexplored at this stage. We mention them here simply in the spirit of identifying potentially interesting directions for future study.

3.11 Conclusion

It is well-established that in a standard model with r factors, the $(r+1)^{th}$ eigenvalue remains bounded for all N. As time goes on, say we encounter a changepoint in which one (or more) of the r factors experiences a change in loadings. In this case, we see that the $(r+1)^{th}$ eigenvalue diverges to infinity with N. This switch in asymptotic behaviour of the $(r+1)^{th}$ eigenvalue precisely at the location of the changepoint implies that a ratio which compares successive values of this eigenvalue over time should spike at the changepoint and remain broadly stable otherwise. In this study, we provide proof of this behaviour and propose that monitoring this ratio over time serves as a reliable basis for sequential changepoint detection. We also prove that this ratio may be estimated from a sample of data consistently as the sample size tends to infinity. We outline a sequential detection strategy which deploys our estimated ratio as a time-varying detection statistic alongside alarm thresholds which we bootstrap using an overlapping blocks resampling scheme. We provide numerical and empirical results to show that our strategy works successfully in practice. We believe our contributions should be useful for any application in which factor models are used to exploit real-time information.

3.12 Appendix – Proofs for Chapter 3

Proof of Lemma 3.4.1

Proof. Let $\psi_j(\cdot)$ denote the jth largest eigenvalue of a matrix.

We recall the result from elementary linear algebra that for a real $n \times m$ matrix M with $m \leq n$, it holds that the m eigenvalues of $m \times m$ matrix $M'M$ are identical to the m non-zero eigenvalues of $n \times n$ matrix MM'. Indeed, it holds that $\psi_j(MM') = 0$ for $m < j \leq n$. See, for instance, Seber (2008, Result 6.54).

In our first case, when $m \leq t \leq \kappa$, under Assumption A, $\Sigma^{\mathbf{c}}_{[m]}(t) = \Gamma AA'\Gamma' = \Gamma A(\Gamma A)' = \Lambda_a \Lambda_a'$, where Λ_a is an $N \times r$ matrix. In our third case, when $\kappa + m - 1 < t$, under Assumption A, $\Sigma^{\mathbf{c}}_{[m]}(t) = \Gamma BB'\Gamma' = \Gamma B(\Gamma B)' = \Lambda_b \Lambda_b'$, where Λ_b is an $N \times r$ matrix. For both these cases, it is immediate from the result quoted above along with Assumption A that there exist constants $\underline{M}_j$ such that

$$\mu^{\mathbf{c}}_{j_{[m]}}(t) \begin{cases} \geq \underline{M}_j N, & j = 1, ..., r \\ = 0, & j = r+1, ..., N. \end{cases}$$

As regards the second case, when $\kappa < t \leq \kappa + m - 1$, under Assumption A,

$$\Sigma^{\mathbf{c}}_{[m]}(t) = \Gamma \left(\frac{m - (t - \kappa)}{m} AA' + \frac{t - \kappa}{m} BB' \right) \Gamma'.$$

Given the definition of matrices A and B, it is clear that

$$AA' = \begin{bmatrix} I_{(r-q)} & 0_{(r-q)\times q} & 0_{(r-q)\times q} \\ 0_{q\times(r-q)} & I_q & 0_{q\times q} \\ 0_{q\times(r-q)} & 0_{q\times q} & 0_{q\times q} \end{bmatrix}, \text{ an } (r+q) \times (r+q) \text{ matrix; and}$$

$$BB' = \begin{bmatrix} I_{(r-q)} & 0_{(r-q)\times q} & 0_{(r-q)\times q} \\ 0_{q\times(r-q)} & 0_{q\times q} & 0_{q\times q} \\ 0_{q\times(r-q)} & 0_{q\times q} & I_q \end{bmatrix}, \text{ an } (r+q) \times (r+q) \text{ matrix.}$$

We observe that both AA' and BB' are positive semidefinite matrices with rank r. On the other hand, the linear combination

$$\frac{m-(t-\kappa)}{m}AA' + \frac{t-\kappa}{m}BB' = \begin{bmatrix} I_{(r-q)} & 0_{(r-q)\times q} & 0_{(r-q)\times q} \\ 0_{q\times(r-q)} & \frac{m-(t-\kappa)}{m}I_q & 0_{q\times q} \\ 0_{q\times(r-q)} & 0_{q\times q} & \frac{t-\kappa}{m}I_q \end{bmatrix},$$

an $(r+q)\times(r+q)$ matrix with only positive entries on the leading diagonal, is positive definite with rank $r+q$.

Let us denote $KK' = \frac{m-(t-\kappa)}{m}AA' + \frac{t-\kappa}{m}BB'$, so that by Merikoski and Kumar (2004, Theorem 7), we may conclude that

$$N\psi_{r+q}\left(KK'\right)\psi_j\left(\frac{\Gamma'\Gamma}{N}\right) \leq \psi_j\left(\Gamma KK'\Gamma'\right) \leq N\psi_1\left(KK'\right)\psi_j\left(\frac{\Gamma'\Gamma}{N}\right)$$

for $j = 1, ..., r+q$. Then, since $\psi_{r+q}\left(KK'\right) = \min\left\{\frac{m-(t-\kappa)}{m}, \frac{t-\kappa}{m}\right\}$, it follows under Assumption A that there exist constants $\underline{M}_j$ such that

$$\mu^{\mathbf{c}}_{j_{[m]}}(t) \begin{cases} \geq \underline{M}_j \min\left\{\frac{m-(t-\kappa)}{m}, \frac{t-\kappa}{m}\right\} N, & j = 1, ..., r+q \\ = 0, & j = r+q+1, ..., N. \end{cases}$$

This completes the proof. □

Proof of Lemma 3.4.2

Proof. Our proof adapts certain techniques from Barigozzi et al. (2021b).

Let $\|\Sigma^{\mathbf{e}}\|_1$, $\|\Sigma^{\mathbf{e}}\|_2$, and $\|\Sigma^{\mathbf{e}}\|_\infty$ denote the column norm, spectral norm, and row norm of $\Sigma^{\mathbf{e}}$ respectively. Then, for any N,

$$\mu^{\mathbf{e}}_{1_{[m]}}(t) = \|\Sigma^{\mathbf{e}}\|_2 \leq \|\Sigma^{\mathbf{e}}\|_1 \leq M \leq \overline{M}_0 < \infty$$

where the first inequality arises due to Hölder's inequality, $\|\Sigma^{\mathbf{e}}\|_2 \leq \sqrt{\|\Sigma^{\mathbf{e}}\|_1\|\Sigma^{\mathbf{e}}\|_\infty}$ along with the recognition that $\|\Sigma^{\mathbf{e}}\|_1$ equals $\|\Sigma^{\mathbf{e}}\|_\infty$, for symmetric matrices, and

the second inequality arises due to Assumption B. The result on the upper bound follows.

The result on the lower bound is a consequence of Assumption B which requires that $\Sigma^{\mathbf{e}}$ is positive definite. Thus, all eigenvalues of $\Sigma^{\mathbf{e}}$ must be positive. □

Proof of Lemma 3.4.3

Proof. Our proof adapts certain techniques from Barigozzi et al. (2021b).

We recall that $\Sigma^{\mathbf{x}}_{[m]}(t) = \Sigma^{\mathbf{c}}_{[m]}(t) + \Sigma^{\mathbf{e}}$ under Assumptions A-C. We apply Weyl's inequality which states that

$$\mu^{\mathbf{c}}_{r+j_{[m]}}(t) + \mu^{\mathbf{e}}_{N_{[m]}}(t) \leq \mu^{\mathbf{x}}_{r+j_{[m]}}(t) \leq \mu^{\mathbf{c}}_{r+j_{[m]}}(t) + \mu^{\mathbf{e}}_{1_{[m]}}(t)$$

for $j = 1, ..., q$.

Consider the cases in which the rolling window consists solely of time periods from the pre- or the post-change regime but not both; that is, cases corresponding either to $m \leq t \leq \kappa$ or to $\kappa + m - 1 < t \leq T$. Then, by combining Weyl's inequality above with Lemmas 3.4.1 and 3.4.2, we know that there exist finite positive constants $\underline{M}_0$ and $\overline{M}_0$ such that

$$\underline{M}_0 \leq \mu^{\mathbf{x}}_{r+j_{[m]}}(t) \leq \overline{M}_0.$$

for $j = 1, ..., q$.

With similar reasoning, for the case in which $\kappa < t \leq \kappa + m - 1$, combining Weyl's inequality with Lemmas 3.4.1 and 3.4.2, we obtain that there exists a finite positive constant $\underline{M}_j$ such that

$$\mu^{\mathbf{x}}_{r+j_{[m]}}(t) \geq \underline{M}_j \min\left\{\frac{t-\kappa}{m}, \frac{\kappa+m-t}{m}\right\} N.$$

for $j = 1, ..., q$. □

Proof of Theorem 3.1

Proof. Using Weyl's inequality, we know that

$$\mu^{\mathbf{c}}_{r+j_{[m]}}(t) + \mu^{\mathbf{e}}_{N_{[m]}}(t) \leq \mu^{\mathbf{x}}_{r+j_{[m]}}(t) \leq \mu^{\mathbf{c}}_{r+j_{[m]}}(t) + \mu^{\mathbf{e}}_{1_{[m]}}(t)$$

for $j = 1, ..., q$. By combining Weyl's inequality above with Lemmas 3.4.1 and 3.4.2, we know that there exist finite positive constants $\underline{M}_0$ and $\overline{M}_0$ such that the denominator of our ratio satisfies

$$\underline{M}_0 \leq \mu^{\mathbf{x}}_{r+j_{[m]}}(m) \leq \overline{M}_0.$$

for $j = 1, ..., q$ since it considers purely the time periods in the pre-change regime (due to Assumption D). On the other hand, the numerator of our ratio, $\mu^{\mathbf{x}}_{r+j_{[m]}}(t)$, behaves exactly as per Lemma 3.4.3. As regards the overall ratio, it follows that

$$\delta_{r+j_{[m]}}(t) = \frac{\mu^{\mathbf{x}}_{r+j_{[m]}}(t)}{\mu^{\mathbf{x}}_{r+j_{[m]}}(m)} \leq \frac{\overline{M}_j}{\underline{M}_0} < \infty$$

if it is the case that $m \leq t \leq \kappa$ or it is the case that $\kappa + m - 1 < t \leq T$, but

$$\delta_{r+j_{[m]}}(t) = \frac{\mu^{\mathbf{x}}_{r+j_{[m]}}(t)}{\mu^{\mathbf{x}}_{r+j_{[m]}}(m)} \geq \min\left\{\frac{t-\kappa}{m}, \frac{\kappa+m-t}{m}\right\}\frac{\underline{M}_j}{\overline{M}_0}N$$

in the case that $\kappa < t \leq \kappa + m - 1$. □

Proof of Lemma 3.5.1

Proof. Our proof adapts certain techniques from Forni et al. (2009).

We consider three cases in which: (i) $m \leq t \leq \kappa$, (ii) $\kappa + m - 1 < t \leq T$, and (iii) $\kappa < t \leq \kappa + m - 1$.

(i) When $m \leq t \leq \kappa$,

$$\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t) = \frac{1}{m} \sum_{h=t-m+1}^{t} x_{ih}x_{jh} = \frac{1}{m} \sum_{h=t-m+1}^{t} x^{(a)}_{ih}x^{(a)}_{jh} = \widehat{\sigma}^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t).$$

We also have that

$$\sigma^{\mathbf{x}}_{ij_{[m]}}(t) = \frac{1}{m} \sum_{h=t-m+1}^{t} \mathsf{E}\left[x_{ih}x_{jh}\right] = \frac{1}{m} \sum_{h=t-m+1}^{t} \mathsf{E}\left[x^{(a)}_{ih}x^{(a)}_{jh}\right] = \sigma^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t).$$

Thus, by Assumption E, $m\mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}}_{ij_{[m]}}(t)\right)^2\right] < \nu_a$.

As previously declared, let the notation $\psi_s(\cdot)$ denote the s^{th} largest eigenvalue of a matrix. We observe that

$$\begin{aligned}
\left\|\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t) - \Sigma^{\mathbf{x}}_{[m]}(t)\right\|_2^2 &= \psi_1\left(\left(\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t) - \Sigma^{\mathbf{x}}_{[m]}(t)\right)\left(\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t) - \Sigma^{\mathbf{x}}_{[m]}(t)\right)'\right) \\
&\leq \mathrm{Tr}\left(\left(\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t) - \Sigma^{\mathbf{x}}_{[m]}(t)\right)\left(\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t) - \Sigma^{\mathbf{x}}_{[m]}(t)\right)'\right) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}}_{ij_{[m]}}(t)\right)^2.
\end{aligned}$$

Taking expectations and using the aforesaid implication of Assumption E,

$$\begin{aligned}
\mathsf{E}\left[\left\|\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t) - \Sigma^{\mathbf{x}}_{[m]}(t)\right\|_2^2\right] &\leq \sum_{i=1}^{N}\sum_{j=1}^{N}\mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}}_{ij_{[m]}}(t)\right)^2\right] \\
&\leq \sum_{i=1}^{N}\sum_{j=1}^{N}\nu_a/m = N^2\nu_a/m.
\end{aligned}$$

Alternatively, we can write

$$\mathsf{E}\left[\left(\frac{1}{N}\left\|\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t) - \Sigma^{\mathbf{x}}_{[m]}(t)\right\|_2\right)^2\right] \leq \nu_a/m.$$

Then, by Chebyshev's inequality, for any positive finite constant c,

$$\Pr\left(\left|\frac{m^{1/2}}{N}\left\|\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t)-\Sigma^{\mathbf{x}}_{[m]}(t)\right\|_2\right|>c\right)\leq \nu_a/c^2.$$

Thus,

$$\frac{1}{N}\left\|\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t)-\Sigma^{\mathbf{x}}_{[m]}(t)\right\|_2=O_p\left(\frac{1}{\sqrt{m}}\right).$$

Then, for the s^{th} eigenvalues $\widehat{\mu}^{\mathbf{x}}_{s_{[m]}}(t)$ and $\mu^{\mathbf{x}}_{s_{[m]}}(t)$ of $\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t)$ and $\Sigma^{\mathbf{x}}_{[m]}(t)$ respectively, we can use Weyl's inequality to show that for $s=1,...,N$,

$$\frac{\widehat{\mu}^{\mathbf{x}}_{s_{[m]}}(t)}{N}\leq\frac{1}{N}\psi_1\left(\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t)-\Sigma^{\mathbf{x}}_{[m]}(t)\right)+\frac{1}{N}\psi_s\left(\Sigma^{\mathbf{x}}_{[m]}(t)\right).$$

Further to this, we obtain

$$\left|\frac{\widehat{\mu}^{\mathbf{x}}_{s_{[m]}}(t)}{N}-\frac{\mu^{\mathbf{x}}_{s_{[m]}}(t)}{N}\right|\leq\frac{1}{N}\left\|\widehat{\Sigma}^{\mathbf{x}}_{[m]}(t)-\Sigma^{\mathbf{x}}_{[m]}(t)\right\|_2=O_p\left(\frac{1}{\sqrt{m}}\right).$$

This completes the proof for the first case.

(ii) When $m\leq t\leq\kappa$,

$$\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t)=\frac{1}{m}\sum_{h=t-m+1}^{t}x_{ih}x_{jh}=\frac{1}{m}\sum_{h=t-m+1}^{t}x^{(b)}_{ih}x^{(b)}_{jh}=\widehat{\sigma}^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t).$$

We also have that

$$\sigma^{\mathbf{x}}_{ij_{[m]}}(t)=\frac{1}{m}\sum_{h=t-m+1}^{t}\mathsf{E}\left[x_{ih}x_{jh}\right]=\frac{1}{m}\sum_{h=t-m+1}^{t}\mathsf{E}\left[x^{(b)}_{ih}x^{(b)}_{jh}\right]=\sigma^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t).$$

Thus, by Assumption E, $m\mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t)-\sigma^{\mathbf{x}}_{ij_{[m]}}(t)\right)^2\right]<\nu_b$. Now, if we follow the proof of the first case above but simply replace ν_a by ν_b therein, we obtain the required result also for the second case.

(iii) When $\kappa < t \leq \kappa + m - 1$,

$$\begin{aligned}
\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t) &= \frac{1}{m} \sum_{h=t-m+1}^{t} x_{ih} x_{jh} \\
&= \frac{\kappa + m - t}{m} \sum_{h=t-m+1}^{t} x^{(a)}_{ih} x^{(a)}_{jh} + \frac{t - \kappa}{m} \sum_{h=t-m+1}^{t} x^{(b)}_{ih} x^{(b)}_{jh} \\
&= \left(\frac{\kappa + m - t}{m} \right) \widehat{\sigma}^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) + \left(\frac{t - \kappa}{m} \right) \widehat{\sigma}^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t).
\end{aligned}$$

We also have that

$$\begin{aligned}
\sigma^{\mathbf{x}}_{ij_{[m]}}(t) &= \frac{1}{m} \sum_{h=t-m+1}^{t} \mathsf{E}\left[x_{ih} x_{jh} \right] \\
&= \left(\frac{\kappa + m - t}{m} \right) \sigma^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) + \left(\frac{t - \kappa}{m} \right) \sigma^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t).
\end{aligned}$$

For convenience, we refer to the weights respectively as w_a and w_b. Hence,

$$\sigma^{\mathbf{x}}_{ij_{[m]}}(t) = w_a \sigma^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) + w_b \sigma^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t),$$

where $w_a, w_b \in (0, 1)$. Then,

$$\begin{aligned}
&\mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}}_{ij_{[m]}}(t) \right)^2 \right] \\
&= \mathsf{E}\left[\left(\left(w_a \widehat{\sigma}^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) + w_b \widehat{\sigma}^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) \right) - \left(w_a \sigma^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) + w_b \sigma^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) \right) \right)^2 \right] \\
&= \mathsf{E}\left[\left(w_a \left(\widehat{\sigma}^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) \right) + w_b \left(\widehat{\sigma}^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) \right) \right)^2 \right] \\
&= w_a^2 \mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) \right)^2 \right] + w_b^2 \mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) \right)^2 \right] \\
&+ 2 w_a w_b \mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}^{(a)}}_{ij_{[m]}}(t) \right) \left(\widehat{\sigma}^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}^{(b)}}_{ij_{[m]}}(t) \right) \right].
\end{aligned}$$

The weights in the above expression, w_a^2, w_b^2, and $w_a w_b$ are bounded from above by 1 for any m. Moreover, if we were to multiply the entire expression by m, then under Assumption E, the first term would be bounded from

above by ν_a, and the second term by ν_b. Additionally, since the first two terms would be bounded from above, we know from the Cauchy-Schwarz inequality that the third term would also be bounded from above. Hence, there exists some positive real ν such that

$$m\mathsf{E}\left[\left(\widehat{\sigma}^{\mathbf{x}}_{ij_{[m]}}(t) - \sigma^{\mathbf{x}}_{ij_{[m]}}(t)\right)^2\right] < \nu.$$

Now, if we follow the proof of the first case above but simply replace ν_a by ν therein, we obtain the required result also for the third case.

This completes the proof. □

Proof of Theorem 3.2

Proof. From Lemma 3.5.1, we have that for any $t \in \{m, ..., T\}$ and $s \in \{1, ..., N\}$,

$$\widehat{\mu}^{\mathbf{x}}_{s_{[m]}}(t) = \mu^{\mathbf{x}}_{s_{[m]}}(t) + O_p\left(N/\sqrt{m}\right).$$

Therefore,

$$\begin{aligned}
\widehat{\delta}_{r+j_{[m]}}(t) &= \frac{\mu^{\mathbf{x}}_{s_{[m]}}(t) + O_p\left(N/\sqrt{m}\right)}{\mu^{\mathbf{x}}_{s_{[m]}}(m) + O_p\left(N/\sqrt{m}\right)} \\
&= \begin{cases} \frac{O(1)+O_p\left(N/\sqrt{m}\right)}{O(1)+O_p\left(N/\sqrt{m}\right)} = O_p(1), & m \leq t \leq \kappa \\ \frac{O(N)+O_p\left(N/\sqrt{m}\right)}{O(1)+O_p\left(N/\sqrt{m}\right)} = O_p(\sqrt{m}), & \kappa < t^*_{N,m} \leq t \leq t^{**}_{N,m} \end{cases}
\end{aligned}$$

and the required result follows.

□

"One can FT anything – often meaningfully."

– John W. Tukey

Chapter 4

Determining Number of Factors via Spectral Eigenvalues

We conduct an exploratory analysis which aims to extend the randomised sequential procedure of Trapani (2018) into the frequency domain. Specifically, we aim to estimate the number of dynamically loaded factors by applying the test of Trapani (2018) to eigenvalues of the estimated spectral density matrix (as opposed to the covariance matrix) of the data.

4.1 Introduction

The aim of our final chapter is relatively modest. That is, to propose an extension of the randomised sequential procedure for determining the number of factors of Trapani (2018). The motivation is as follows.

Trapani (2018) considers a static approximate model of the type considered, for example, by Bai and Ng (2002) and Stock and Watson (2002a,b). In this setup, the common component for i^{th} observation at the t^{th} time period, say χ_{it} for $i = 1, ..., N$ and $t = 1, ..., T$, is formed as a linear combination

$$\chi_{it} = \sum_{j=1}^{k} \phi_{ij} f_{jt}$$

of $k \ll N$ latent factors $f_{1t}, ..., f_{kt}$ that are all loaded contemporaneously.

In order to allow for dynamics, one approach might be to attempt to extend the Trapani (2018) procedure to the restricted approximate dynamic model case as espoused *inter alia* by Forni et al. (2009).

Alternatively, as outlined in Chapter 1, we can consider the work say by Hallin and Lippi (2013), which establishes that the approximate dynamic factor model is the only one supported by a representation theorem. Thus, another promising avenue of research would be to mirror the approach of Trapani (2018) in the frequency domain, with the goal of developing a method for direct use in a slightly more general framework, one in which factors are loaded with lags.

Such an exercise would fit well into the literature. Indeed, our aim is to complement the work of Trapani (2018) in exactly the manner that Hallin and Liška (2007) complements Bai and Ng (2002) and Alessi et al. (2010), or that Onatski (2009) complements Ahn and Horenstein (2013).

In this chapter, we recapitulate the steps of the procedure of Trapani (2018) before embarking upon a short numerical foray to explore the feasibility of such an extension into the frequency domain. We draw upon insights obtained from our Chapter 2 where possible. We also outline possible avenues of future research.

4.2 Individual Test

The individual[1] test of Trapani (2018) (hereafter, just "the test") is designed for a hypothesis concerning the p-th largest eigenvalue of the population covariance matrix of a vector-valued second-order stationary time series $\mathbf{x}_t$ that is governed by a static approximate factor model. It is based on the insight that in a k-factor model, the k leading eigenvalues will diverge (linearly with N under typical assumptions) as $N \to \infty$, while the remainder stay bounded. Letting the p-th largest eigenvalue of $\mathsf{E}[\mathbf{x}_t \mathbf{x}_t']$, for $p = 1, ..., N$, be denoted by $\lambda^{(p)}$, we have

$$
\begin{aligned}
H_0 &: \lambda^{(p)} = m_p N \\
H_1 &: \lambda^{(p)} = m_p < \infty,
\end{aligned}
$$

for some finite positive constant m_p.

It is worth stating at this point that we find no benefit to modifying the notation of Trapani (2018) when describing the procedure. Our goal is simply to the recapitulate key steps here for ease of reference and thus we prefer to retain the clarity of the original exposition. Thus, with full credit to the author, we simply borrow the parts needed for our purposes from the original source.

As noted in Trapani (2018, Section 3.1, p.1344):

> Let $\beta \equiv \frac{\ln N}{\ln T}$, and define $\delta \in [0, 1)$ such that
>
> $$
> \delta \begin{cases} > 0 \\ > 1 - \frac{1}{2\beta} \end{cases} \text{ according as } \begin{matrix} \beta \leq \frac{1}{2} \\ \beta > \frac{1}{2} \end{matrix} .
> $$
>
> Finally, consider the following estimator of $\bar{\lambda}_N$
>
> $$
> \widehat{\bar{\lambda}}_N \equiv \frac{1}{N} \sum_{p=1}^{N} \widehat{\lambda}^{(p)}
> $$

[1] "Individual" in the sense that it concerns a test for a specific eigenvalue. Of course, this is but a building block of the fuller analysis in which Trapani (2018) goes on to define a sequential procedure to determine the answer to the ultimate question of how many factors are present.

We are now ready to introduce the test. Define

$$\varphi^{(p)} \equiv \exp\left\{ N^{-\delta} \frac{\widehat{\lambda}^{(p)}}{\widehat{\bar{\lambda}}_N} \right\}$$

Above, the symbol "≡" denotes a definitional equality, and a hat denotes an estimator. $\bar{\lambda}_N$ refers to the mean of $\lambda^{(p)}$ for $p = 1, ..., N$.

Trapani (2018, Section 3.1, p.1344) continues as follows:

> Step 1. Generate an artificial sample $\left\{\xi_j^{(p)}\right\}_{j=1}^{R}$ as iid $\mathcal{N}(0,1)$, and define the sequence $\sqrt{\varphi^{(p)}} \times \xi_j^{(p)}, 1 \leq j \leq R$;
>
> Step 2. Define the sample $\left\{\zeta_j^{(p)}(u)\right\}_{j=1}^{R}$ as
>
> $$\zeta_j^{(p)}(u) \equiv I\left[\sqrt{\varphi^{(p)}} \times \xi_j^{(p)} \leq u\right],$$
>
> with u extracted from a distribution $F(u)$ with support $U \subset R\backslash\{0\}$;
>
> Step 3. Compute
>
> $$\vartheta^{(p)}(u) \equiv \frac{2}{\sqrt{R}} \sum_{j=1}^{R} \left[\zeta_j^{(p)}(u) - \frac{1}{2}\right];$$
>
> Step 4. Define the test statistic
>
> $$\Theta^{(p)} \equiv \int_U \left[\vartheta^{(p)}(u)\right]^2 dF(u).$$

While we were unable to verify its definition, we believe the notation $I[\cdot]$ above refers to an indicator function.

We close this section by noting the key result that $\Theta^{(p)}$ has a chi-squared distribution with one degree of freedom and this forms the basis for the test. (Trapani, 2018, Theorem 3.1)

4.3 Preliminary Findings

We consider a model that introduces dynamics yet remains simple enough so as not to suffer a loss in tractability. In particular, we consider the exact same setup used in the numerical study in Section 2.9 of Chapter 2 (and thus the specification outline is not repeated here). In particular, given a model where the factors are loaded with a single-lag, we begin by setting $N = 100$ and $T = 400$.

We first estimate the $N \times N$ spectral density matrix of $\mathbf{x}_t$ using the lag-window method (with a Bartlett lag-window characterised by truncation point B_T) as per Section 2.6.1 in our Chapter 2. We denote our estimator by $\widetilde{G}^{B_T}_{\mathbf{x}_t}(w_{j'})$ and compute it on a grid of $2B_T+1$ points. We set B_T to $2\lfloor\sqrt{T}\rfloor$, and so with $T = 400$, we obtain $B_T = 40$.[2]

We subsequently obtain dynamic eigenvalues denoted by $\widetilde{\Lambda}_{\mathbf{x}_t}(w_{j'})$ again on a grid of $2B_T+1 = 81$ points via an eigendecomposition of the estimated spectral density matrix. We note that dynamic eigenvalues are real (due to the Hermitian nature of the spectral density).

We present the following charts to illustrate that the behaviour of estimated dynamic eigenvalues in our simulations is as expected. We see evidence of divergence of the first 1 and 2 dynamic eigenvalues in specifications with 1 and 2 dynamic factors respectively. This is compatible with results on uniform divergence across frequencies (linearly with N), for instance, in Hallin and Liška (2007).

We observe that the divergence behaviour appears more pronounced at lower frequencies (but not necessarily in a strictly ordered way). It is worth exploring whether this is a systematic finding we could make use of or simply a result of the model specification being one which is not dominated by high-frequency movements. For now, we simply compute an average of the dynamic eigenvalues (as does Hallin and Liška (2007)) over the frequencies and proceed towards implementation of the test.

[2]From a theoretical perspective, we recall that we require $B_T \to \infty$ and $B_T/T \to 0$ as $T \to \infty$. In numerical studies (where T is given), the choice of B_T is, of course, arbitrary. During our in-depth numerical studies of Chapter 2, we found that setting B_T to $2\lfloor\sqrt{T}\rfloor$ worked well in practice at least in the context of the (N, T) combinations considered therein.

Figure 4.1: Dynamic Eigenvalues by Frequency (1-factor model)

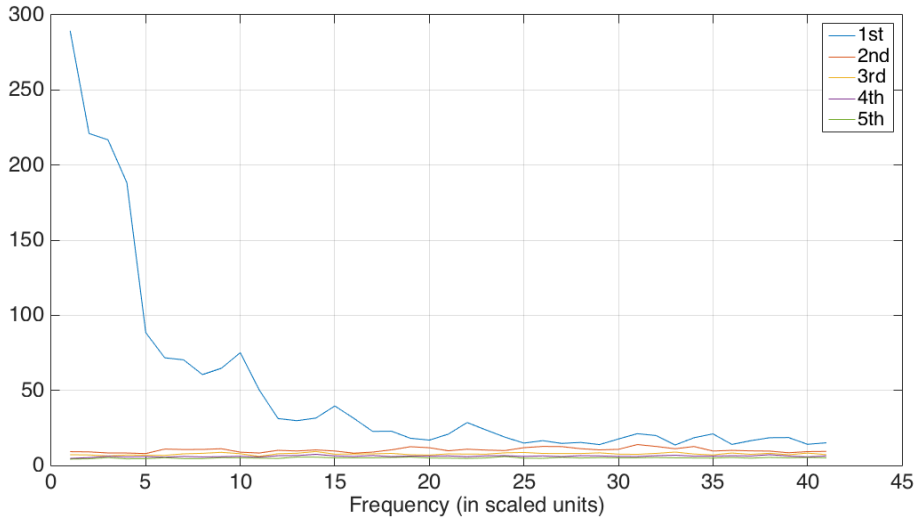
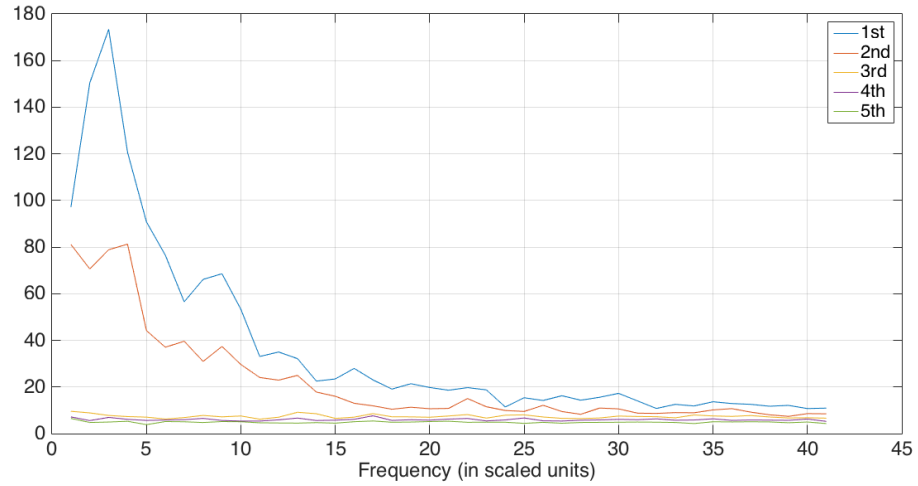


Figure 4.2: Dynamic Eigenvalues by Frequency (2-factor model)



By way of an initial jaunt, we consider the two-factor case under the same settings as above and implement 100 individual tests at the 5% level for the first three eigenvalues.[3] The idea (albeit imprecise) is that the tests for the two leading

[3]For avoidance of doubt, we note that each of the aforesaid tests is based on a sample that is independently drawn. It is not the same sample on which we keep testing multiple eigenvalues.

eigenvalues (wherein the null of divergence is true) should give us a sense of size and the tests for the third eigenvalue (wherein the null is false) should give us a sense of power. Indeed, we find that in the case with no cross-sectional dependence and no serial correlation in the idiosyncratic component, the tests for the first and second eigenvalues are each rejected 6 times out 100; whereas the test for the third eigenvalue is rejected 100 times out of 100. This appears promising and motivates the brief numerical exercise presented below.

In order to investigate size, we compute empirical rejection rates for the null that the 2nd dynamic eigenvalue (averaged over the frequencies) in a two-factor model diverges to infinity. In other words, the null is true and we check how often it is rejected. In order to investigate power, we compute empirical rejection rates for the same null but with a one-factor model. In other words, the null is false and we check how often it is rejected.

Tests are conducted for an illustrative variety of (N, T) combinations. The notation $\widetilde{T}$ refers to the effective sample size of $2B_T + 1$. Each rejection rate reported in the tables below is based on 300 separately (and independently) drawn samples from a synthetic data-generating process as described in detail in Section 2.9 of Chapter 2. All tests are carried out at the 5% level.

Its simplicity notwithstanding, our exercise demonstrates that the procedure appears to work. In other words, estimates of size and power are in line with what we would expect and patterns are not dissimilar to those presented in Trapani (2015, p.28–29), an earlier version of the paper under consideration. Further research in this area definitely appears warranted. We conduct a more in-depth numerical investigation in the subsequent section.

Table 4.1: Empirical Rejection Rates
(no cross-sectional or serial dependence)

k=2					
	T	$\widetilde{T}$		N	
			50	75	100
	100	41	0.053	0.060	0.050
	500	89	0.077	0.057	0.047
	1000	125	0.090	0.057	0.063
	2000	177	0.080	0.057	0.057
k=1					
	T	$\widetilde{T}$		N	
			50	75	100
	100	41	0.973	1.000	1.000
	500	89	0.980	1.000	1.000
	1000	125	0.980	1.000	1.000
	2000	177	0.967	1.000	1.000

Table 4.2: Empirical Rejection Rates
(cross-sectional but not serial dependence)

k=2					
	T	$\widetilde{T}$		N	
			50	75	100
	100	41	0.070	0.070	0.057
	500	89	0.080	0.067	0.043
	1000	125	0.080	0.067	0.053
	2000	177	0.047	0.077	0.070
k=1					
	T	$\widetilde{T}$		N	
			50	75	100
	100	41	0.973	1.000	1.000
	500	89	0.970	1.000	1.000
	1000	125	0.980	1.000	1.000
	2000	177	0.917	0.993	1.000

Table 4.3: Empirical Rejection Rates
(cross-sectional and serial dependence)

k=2					
	T	$\widetilde{T}$		N	
			50	75	100
	100	41	0.067	0.047	0.050
	500	89	0.050	0.087	0.073
	1000	125	0.050	0.037	0.063
	2000	177	0.050	0.093	0.043
k=1					
	T	$\widetilde{T}$		N	
			50	75	100
	100	41	0.573	0.753	0.877
	500	89	0.400	0.653	0.860
	1000	125	0.277	0.640	0.870
	2000	177	0.177	0.537	0.877

4.4 Determining the Number of Factors

The individual tests considered thus far can be cast in a procedure to determine the number of common factors in a dynamic approximate factor model. In this section, we first explain this procedure and then conduct a simulation study to evaluate its performance under a variety of different data-generating processes. Hereafter, we refer to the procedure as the spectral version of the randomised sequential procedure of Trapani (2018).

4.4.1 Spectral Randomised Sequential Procedure

The individual tests can be employed in a sequential manner to determine the number of dynamic factors by the following two-step approach described in Trapani (2018, p.1345):

Step 1. We run a test for divergence of the largest dynamic eigenvalue based on test statistic $\Theta^{(1)}$ as previously described. If the null is rejected, we set our estimator of the number of dynamic factors to be $\widehat{k} = 0$ and terminate the procedure; otherwise, we go to the next step.

Step 2. Commencing with $p = 1$, we test for divergence of the $(p+1)^{th}$ dynamic eigenvalue based on test statistic $\Theta^{(p+1)}$, constructed using an artificial sample $\left\{\xi_j^{(p+1)}\right\}_{j=1}^{R}$, generated independently of $\left\{\xi_j^{(1)}\right\}_{j=1}^{R}, ..., \left\{\xi_j^{(p)}\right\}_{j=1}^{R}$. If the null is rejected, we set $\widehat{k} = p$ and terminate the procedure; otherwise, we repeat this step until the null is rejected (or until a pre-specified maximum, say k_{max}, is reached).

4.4.2 Monte Carlo Study

We now report the results of experiments that directly address the question of what is the number of factors. Experiments are carried out for a variety of (N, T) combinations and data-generating processes. In particular, we examine estimates of the number of factors over simulations repeated 100 times. Again, the synthetic data-generating process is exactly as described in Section 2.9 of Chapter 2. All tests are carried out at the 5% level.[4]

With respect to control parameters, we let cross-section $N \in \{25, 50, 75, 100, 125\}$, time periods $T \in \{100, 200, 500, 1000\}$, number of factors $k \in \{1, 2\}$, noise-signal ratio $\rho \in \left\{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}\right\}$, cross-sectional dependence $\phi \in \left\{0, \frac{1}{3}, \frac{2}{3}\right\}$, and serial dependence $b \in \left\{0, \frac{1}{3}, \frac{2}{3}\right\}$. We also examine the effect of dependence arising due to the law of motion for the factors, $a \in \left\{0, \frac{1}{3}, \frac{2}{3}\right\}$. Spectral density estimates and calculation of dynamic eigenvalues via dynamic PCA (Brillinger, 1981, Chapter 9) are exactly as described in Chapter 2, and k_{max} is set to 4.

We make available the full set of tables from our Monte Carlo study in Appendix 4.6 below. The tables contain for each (N, T) configuration: (i) the average estimated number of factors, $\hat{k}$, and (ii) the percentage of incorrect predictions

[4]Following the recommendation in Trapani (2018), we also employ a Bonferroni-style correction. See Trapani (2018, p.1346); also see Trapani (2015, p.19) for further details.

– i.e. where $\hat{k} \neq k$. We refer to (i) as "Av.$\hat{k}$", and to (ii) as the "Rate of Error (RoE)". We present each of these two summary statistics under low, moderate and reasonably-high noise settings (see values for parameter ρ above).

Our main findings, along with cross-references to relevant tables, are as follows:

(i) We observe excellent performance in both the 1 and 2 factor cases, even under moderate and high noise settings, provided that N and T are of sufficient size. In particular, results vastly improve for $N \geq 75$ and $\widetilde{T} \geq 89$. Consider, for instance, Tables 4.4 and 4.10. (This is typical under the usual double-asymptotic framework considered in the literature.)

(ii) There is a very slight deterioration in performance for the 2 factor case (Table 4.10) in comparison with the 1 factor case (Table 4.4) in the sense that we need slightly higher sample sizes (e.g. $\widetilde{T} \geq 125$) for estimates to stabilise at the true value of k. (This is to be expected given that the magnitude of the second largest eigenvalue is, trivially speaking, smaller than that of the largest eigenvalue; therefore, the signal gets weaker.)

(iii) The procedure appears remarkably capable of handling cross-sectional dependence in the idiosyncratic component – see Tables 4.5–4.6 (and Tables 4.11–4.12). Results do deteriorate relative to the baseline Table 4.4 (and Table 4.10) discussed under point (i) above but good performance is maintained so long as sample sizes remain reasonably high as previously described. (As regards direction, the tendency – in almost all configurations – is for us to over-estimate the number of factors when cross-sectional dependence is introduced. We believe this is in accordance with expectations.)

(iv) In contrast with point (iii) above, serial dependence in the idiosyncratic component appears more concerning. Indeed, moderate serial dependence brings about a notable deterioration in performance – see Table 4.7 (and Table 4.13) – particularly for low values of N and T. Further, high serial dependence – see Table 4.8 (and Table 4.14) – systematically forces over-estimation of the number of dynamic factors. We remark that this finding appears in accordance with the motivation put forward by Lam and Yao (2012) for their procedure. Indeed, as explained therein,

> An idiosyncratic series may exhibit serial correlations and, therefore, may be a time series itself. This poses technical difficulties in both model identification and inference. In fact the rigorous definition of the common factors and the idiosyncratic components can only be established asymptotically when the dimension of time series tends to infinity. (Lam and Yao, 2012, p.696)

(Interestingly, Trapani (2018, p.1346, p.1348) hints that it is cross-sectional dependence that affects simulation results therein relatively more than serial dependence. Our conjecture is that the reversal in pattern is likely to be an artefact of the fact that we are working in the frequency domain, but further investigation would be needed to confirm the mechanism by which this is borne out.)

(v) Given the above finding, we chose to consider the impact of an alternative source of dynamics in observables. That is, we model a law of motion for the dynamically loaded factors themselves – see Tables 4.16–4.17. Again, there is a slight deterioration in the performance of the estimator but results are much improved in comparison with point (iv) above.

4.4.3 Empirical Illustration

We now apply the spectral randomised sequential procedure to a widely-studied real-world dataset. Specifically, we consider a dataset on 112 monthly US macroeconomic series between March 1973 and November 2007 (for a total of $T = 416$ data points) that was carefully scrutinised by Forni and Gambetti (2010). We refer the reader to Forni and Gambetti (2010, Appendix A. Supplementary Material) for a precise list of the series covered (and any transformations applied thereto).

As mentioned in Chapter 1, the paper is well-known for solving many so-called "puzzles" previously unaccounted for by methodologies based on structural vector autoregressive (SVAR) models in empirical macroeconometrics. Indeed, Forni and Gambetti (2010) analyses the effects of monetary policy shocks under a

factor model (Forni et al., 2009) identified using a standard recursive scheme in which the impact effects on both industrial production and prices are zero. The findings are that: (i) the maximal effect on bilateral real exchange rates is observed on impact, so that the "delayed overshooting" or "forward discount" puzzle disappears; (ii) after a contractionary shock prices fall at all horizons, so that the price puzzle is not there; and (iii) monetary policy has a sizeable effect on both real and nominal variables.

The dataset – or more precisely, an older version thereof – has also been studied by Bernanke et al. (2005), and prior to that by Stock and Watson (1999). The dataset we use – that is, the version of Forni and Gambetti (2010) – may be found on the *Journal of Monetary Economics* website, whereas the prior version may be found on the *Princeton University* website[5] maintained by Professor Watson.

Following the imposition of a factor model structure, one of the first aims in Forni and Gambetti (2010) is to ascertain the number of dynamic factors (or structural shocks). The paper states:

> To determine the number of dynamic factors, we used the criteria proposed by Bai and Ng (2007), Amengual and Watson (2007) and Hallin and Liska (2007). (Forni and Gambetti, 2010, Section 3.2)

The paper concludes:

> [...] that the number of dynamic factors is in the interval 4–7. In our preferred model *we use 4 dynamic factors* [...]. (Forni and Gambetti, 2010, Section 3.2, emphasis added)

In this empirical example, we try to match the findings above using two procedures: the first is that of Hallin and Liška (2007), and the second is our spectral randomised sequential procedure. Our overall aim is simply to verify that the various methodologies considered are able to corroborate each other.

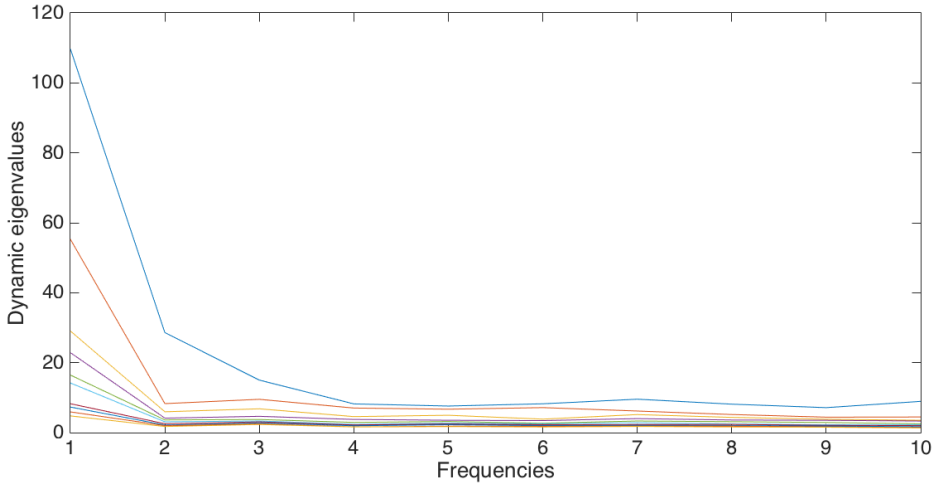
Figure 4.3 presents estimated dynamic eigenvalues by frequency. There is no clear eigen-gap immediately discernible. Nevertheless, it could be argued that a visual inspection might suggest up to around 6 dynamic factors.

[5] https://www.princeton.edu/~mwatson/publi.html

Figure 4.4 is the plot obtained by applying the method[6] of Hallin and Liška (2007). We note that c is a tuning parameter that determines the intensity with which the penalty associated with over-fitting takes effect. To interpret Figure 4.4, we appeal to the theoretical arguments given by Hallin and Liška (2007). The theory suggests that the relevant interval to consider is the second stability interval, corresponding to the smallest values of c for which the estimator for k is constant. (The reason the first stability interval is excluded is because it always corresponds to the boundary solution k_{max} and is thereby deemed non-admissible.) Based on Figure 4.4, one could argue that there are around 3 to 6 dynamic factors present.

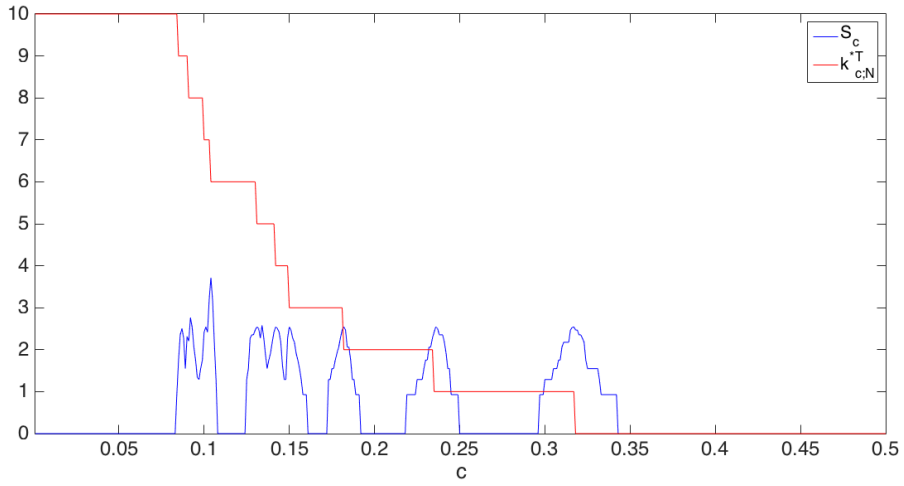
Finally, we apply the spectral randomised sequential procedure using the same configuration as per the previous section, with the exception that k_{max} is increased to 10 in this case. Interestingly, we obtained a result of exactly 4 dynamic factors, which was precisely the conjecture of Forni and Gambetti (2010) mentioned above. We do not read too much into this result other than to remark that it does seem to be pointing in the right direction.

Figure 4.3: Dynamic Eigenvalues by Frequency (US macro data)



[6]We use the so-called "$IC^T_{2;N}(k)$" criterion.

Figure 4.4: Criterion of Hallin and Liška (2007) – Summary Plot



4.5 Future Directions

This chapter is, of course, the subject of ongoing work and there is a lot that could be done. To reiterate, we believe that (i) the proposed premise (for conducting research in this area) is sensible; and (ii) our rudimentary numerical investigation above supports the premise.

As regards immediate next steps, our plan is to carry out a much more comprehensive numerical study. For instance, we could expand the N, T combinations considered, and also vary the modelling structure in different ways. Interesting directions could include investigating the impact of: (i) different dynamical structures (i.e. increasing the number of lags with which factors are loaded, varying the magnitude of serial correlation in the factors and idiosyncratic components, attempting more complex ARMA specifications); (ii) different sizes of B_T and/or potentially different estimators of the spectral density; (iii) adjusting the exogenous parameters[7] such as R, δ, and the support for u; and (iv) changing the distributional assumptions for the data-generating process (e.g. heavy-tails). As-

[7] In the numerical exercises presented above, we had simply set these parameters in accordance with suggestions in Trapani (2015, 2018)

suming we obtain a series of sensible results in the simulation studies, a priority would be to test the procedure on a real macroeconomic dataset and compare it with the performance of competing estimators mentioned previously. Subject to the results of the proposed investigative empirical analysis, we plan to then initiate a comprehensive examination of the theory via leveraging results from Trapani (2018) where possible.

Moreover, let us suppose, with a dose of healthy optimism, that we are successfully able to develop the theory for a spectral version of the test. It would then be interesting to consider an extension in respect of the case, say, in which a (restricted approximate dynamic) model consists of an r-dimensional vector of $I(1)$ factors that are driven by a smaller number q of common shocks. In this setting, Barigozzi et al. (2021b) explains that the factors are cointegrated and discusses why this is a macroeconomically relevant situation. In particular, the framework therein is that we have r factors and q common shocks, of which τ are permanent (e.g. technological) and d (e.g. monetary policy) are transitory. Interestingly for us, Barigozzi et al. (2021b, Section 4, p.465) states:

> It can be shown that [...] the q largest eigenvalues of [the spectral density of $\mathbf{x}_t$ at frequency $\theta \in [-\pi, \pi)$] diverge linearly in $[N]$, while the remaining $[N - q]$ stay bounded. This is true at all frequencies but at frequency $\theta = 0$, where only the τ largest eigenvalues of [the spectral density at frequency 0] diverge linearly in N.

In other words, it is theoretically possible to determine the values of q and τ by analysing the behaviour of the eigenvalues of the spectral density matrix at the appropriate frequencies. Barigozzi et al. (2021b, Section 4) considers a very useful extension in this respect to the Hallin and Liška (2007) and Onatski (2009) approaches. Correspondingly, our plan is to leverage any theoretical results that emanate from our own work proposed in this chapter to also consider specifically the zero-frequency case.

4.6 Appendix – Full Set of Simulation Results

We gather, in this Appendix, the full set of simulation results from our Monte Carlo study on determining the number of factors. Given the large number of tables provided, we present here a summary of the contents of this section.

Tables 4.4–4.9 consider the 1 factor setting whereby:

- Table 4.4 considers absence of any cross-sectional or serial dependence;
- Tables 4.5–4.6 consider presence of cross-sectional dependence only;
- Tables 4.7–4.8 consider presence of serial dependence only;
- Table 4.9 considers presence of both cross-sectional and serial dependence.

Tables 4.10–4.15 consider analogous results for the 2 factor setting:

- Table 4.10 considers absence of any cross-sectional or serial dependence;
- Tables 4.11–4.12 consider presence of cross-sectional dependence only;
- Tables 4.13–4.14 consider presence of serial dependence only;
- Table 4.15 considers presence of both cross-sectional and serial dependence.

Tables 4.16–4.17 consider different laws of motion for the factor.

Table 4.4: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 1, \phi = 0, b = 0$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	2.08	78%	2.59	92%	2.71	95%
100	41	50	1.24	23%	1.50	46%	1.71	56%
100	41	75	1.00	2%	1.09	15%	1.17	24%
100	41	100	0.97	3%	0.99	3%	1.03	7%
100	41	125	0.94	6%	0.96	6%	0.93	7%
500	89	25	2.56	88%	2.90	96%	3.37	98%
500	89	50	1.28	27%	1.67	55%	2.01	71%
500	89	75	1.03	3%	1.08	8%	1.18	18%
500	89	100	1.00	0%	1.02	0%	1.01	1%
500	89	125	1.00	0%	1.00	2%	1.00	0%
1000	125	25	2.91	93%	3.43	99%	3.62	99%
1000	125	50	1.37	34%	1.65	51%	2.01	69%
1000	125	75	1.01	1%	1.09	9%	1.24	24%
1000	125	100	1.00	0%	1.01	1%	1.02	2%
1000	125	125	1.00	0%	1.00	0%	1.00	0%
2000	177	25	3.22	94%	3.63	99%	3.75	100%
2000	177	50	1.39	36%	1.74	60%	2.38	87%
2000	177	75	1.04	4%	1.18	18%	1.31	28%
2000	177	100	1.00	0%	1.03	3%	1.02	2%
2000	177	125	1.00	0%	1.00	0%	1.01	1%

Table 4.5: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 1, \phi = \frac{1}{3}, b = 0$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	2.12	79%	2.36	88%	2.83	95%
100	41	50	1.23	25%	1.40	39%	1.66	61%
100	41	75	1.02	2%	1.12	14%	1.17	23%
100	41	100	1.00	2%	0.97	3%	1.04	8%
100	41	125	0.95	5%	0.98	2%	0.99	5%
500	89	25	2.70	93%	3.23	98%	3.54	99%
500	89	50	1.25	24%	1.58	54%	1.91	71%
500	89	75	1.04	4%	1.05	5%	1.20	18%
500	89	100	1.00	0%	1.01	1%	1.00	2%
500	89	125	1.00	0%	0.99	1%	1.00	0%
1000	125	25	2.81	91%	3.30	95%	3.62	99%
1000	125	50	1.34	33%	1.69	58%	2.09	78%
1000	125	75	1.01	1%	1.07	6%	1.24	24%
1000	125	100	1.00	0%	1.00	0%	1.03	3%
1000	125	125	1.00	0%	1.00	0%	1.01	1%
2000	177	25	3.36	99%	3.65	100%	3.74	100%
2000	177	50	1.54	44%	1.84	62%	2.42	89%
2000	177	75	1.04	4%	1.13	15%	1.35	34%
2000	177	100	1.00	0%	1.01	1%	1.03	3%
2000	177	125	1.00	0%	1.00	0%	1.00	0%

Table 4.6: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 1, \phi = \frac{2}{3}, b = 0$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	2.10	82%	2.32	87%	2.87	97%
100	41	50	1.32	34%	1.40	38%	1.74	60%
100	41	75	1.01	3%	1.13	15%	1.22	24%
100	41	100	1.00	0%	1.00	2%	1.08	14%
100	41	125	0.97	3%	0.96	4%	0.97	9%
500	89	25	2.86	97%	3.16	99%	3.36	99%
500	89	50	1.38	37%	1.63	52%	1.98	76%
500	89	75	1.04	4%	1.09	9%	1.31	29%
500	89	100	1.01	1%	1.00	0%	1.01	1%
500	89	125	1.00	0%	1.01	1%	1.00	0%
1000	125	25	2.93	95%	3.42	100%	3.69	100%
1000	125	50	1.42	38%	1.93	67%	2.35	88%
1000	125	75	1.07	7%	1.17	16%	1.50	42%
1000	125	100	1.00	0%	1.02	2%	1.07	7%
1000	125	125	1.00	0%	1.00	0%	1.01	1%
2000	177	25	3.28	98%	3.69	99%	3.83	100%
2000	177	50	1.60	55%	2.13	74%	2.57	89%
2000	177	75	1.07	7%	1.28	26%	1.54	50%
2000	177	100	1.00	0%	1.00	0%	1.07	7%
2000	177	125	1.00	0%	1.01	1%	1.02	2%

Table 4.7: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 1, \phi = 0, b = \frac{1}{3}$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	2.71	91%	3.18	97%	3.28	100%
100	41	50	1.84	67%	2.38	92%	2.72	96%
100	41	75	1.23	23%	1.69	61%	2.07	83%
100	41	100	1.10	12%	1.43	40%	1.75	69%
100	41	125	1.06	8%	1.24	27%	1.45	44%
500	89	25	3.58	100%	3.79	100%	3.85	100%
500	89	50	2.08	82%	2.91	96%	3.47	98%
500	89	75	1.39	36%	2.02	79%	2.40	87%
500	89	100	1.05	5%	1.43	39%	1.89	64%
500	89	125	1.02	2%	1.15	17%	1.41	39%
1000	125	25	3.72	100%	3.86	100%	3.95	100%
1000	125	50	2.30	82%	3.11	96%	3.64	99%
1000	125	75	1.52	48%	1.99	79%	2.81	94%
1000	125	100	1.11	11%	1.42	40%	1.97	74%
1000	125	125	1.01	1%	1.07	7%	1.38	37%
2000	177	25	3.91	100%	3.98	100%	4.00	100%
2000	177	50	2.76	94%	3.42	98%	3.84	99%
2000	177	75	1.52	48%	2.26	83%	2.92	93%
2000	177	100	1.12	12%	1.38	35%	1.96	72%
2000	177	125	1.02	2%	1.12	12%	1.41	40%

Table 4.8: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 1, \phi = 0, b = \frac{2}{3}$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	3.56	100%	3.81	100%	3.89	100%
100	41	50	3.08	98%	3.52	100%	3.78	100%
100	41	75	2.35	91%	3.15	100%	3.52	100%
100	41	100	2.11	89%	2.84	99%	3.21	100%
100	41	125	1.83	69%	2.49	95%	2.96	99%
500	89	25	3.97	100%	3.98	100%	3.97	99%
500	89	50	3.76	100%	3.93	100%	3.97	99%
500	89	75	2.98	96%	3.75	99%	3.98	100%
500	89	100	2.42	86%	3.28	99%	3.82	100%
500	89	125	1.79	71%	2.77	94%	3.67	100%
1000	125	25	3.97	100%	4.00	100%	4.00	100%
1000	125	50	3.78	100%	3.98	100%	4.00	100%
1000	125	75	3.09	99%	3.89	100%	4.00	100%
1000	125	100	2.49	89%	3.62	99%	3.97	100%
1000	125	125	1.98	77%	3.02	98%	3.67	100%
2000	177	25	4.00	100%	4.00	100%	3.97	99%
2000	177	50	3.89	100%	3.99	100%	4.00	100%
2000	177	75	3.60	100%	3.99	100%	3.98	100%
2000	177	100	2.61	92%	3.82	99%	3.93	100%
2000	177	125	2.10	79%	3.25	99%	3.84	99%

Table 4.9: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 1, \phi = \frac{1}{3}, b = \frac{1}{3}$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	2.88	98%	3.27	100%	3.48	100%
100	41	50	1.85	71%	2.38	91%	2.63	94%
100	41	75	1.28	28%	1.82	67%	2.18	84%
100	41	100	1.10	14%	1.45	46%	1.80	66%
100	41	125	0.99	7%	1.17	19%	1.42	40%
500	89	25	3.60	100%	3.79	99%	3.87	100%
500	89	50	2.19	84%	2.84	94%	3.40	100%
500	89	75	1.37	36%	2.05	77%	2.54	92%
500	89	100	1.07	7%	1.33	32%	1.95	75%
500	89	125	1.01	1%	1.12	12%	1.42	40%
1000	125	25	3.67	100%	3.94	100%	3.97	100%
1000	125	50	2.45	88%	3.11	95%	3.62	99%
1000	125	75	1.55	51%	2.18	78%	2.89	96%
1000	125	100	1.02	2%	1.50	46%	1.87	69%
1000	125	125	1.01	1%	1.18	18%	1.36	34%
2000	177	25	3.92	100%	3.98	100%	4.00	100%
2000	177	50	2.69	94%	3.55	99%	3.81	100%
2000	177	75	1.54	51%	2.38	85%	2.94	94%
2000	177	100	1.12	12%	1.53	50%	2.11	76%
2000	177	125	1.01	1%	1.22	21%	1.58	49%

Table 4.10: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 2, \phi = 0, b = 0$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	3.11	76%	3.41	88%	3.49	92%
100	41	50	1.99	21%	2.13	32%	2.32	41%
100	41	75	1.64	35%	1.74	30%	1.76	37%
100	41	100	1.51	49%	1.55	43%	1.65	36%
100	41	125	1.36	62%	1.47	53%	1.45	52%
500	89	25	3.54	90%	3.75	97%	3.85	96%
500	89	50	2.09	9%	2.37	32%	2.59	52%
500	89	75	2.00	0%	2.00	4%	2.01	7%
500	89	100	1.99	1%	1.99	1%	1.94	6%
500	89	125	1.98	2%	1.91	9%	1.87	13%
1000	125	25	3.64	90%	3.84	98%	3.88	98%
1000	125	50	2.12	10%	2.33	32%	2.75	59%
1000	125	75	2.00	0%	2.01	1%	2.06	6%
1000	125	100	2.00	0%	2.00	0%	2.00	0%
1000	125	125	2.00	0%	2.00	0%	1.98	2%
2000	177	25	3.78	99%	3.86	98%	3.98	100%
2000	177	50	2.22	21%	2.47	40%	2.70	56%
2000	177	75	2.00	0%	2.02	2%	2.07	7%
2000	177	100	2.00	0%	2.00	0%	2.00	0%
2000	177	125	2.00	0%	2.00	0%	2.00	0%

Table 4.11: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 2, \phi = \frac{1}{3}, b = 0$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	3.05	75%	3.25	84%	3.43	89%
100	41	50	1.96	16%	2.16	32%	2.41	50%
100	41	75	1.68	31%	1.75	27%	1.86	24%
100	41	100	1.65	34%	1.58	41%	1.59	42%
100	41	125	1.59	39%	1.51	47%	1.48	49%
500	89	25	3.59	96%	3.72	94%	3.86	100%
500	89	50	2.10	10%	2.35	31%	2.55	46%
500	89	75	1.99	1%	1.99	3%	2.07	7%
500	89	100	1.98	2%	1.98	2%	1.99	1%
500	89	125	1.99	1%	1.90	10%	1.89	11%
1000	125	25	3.63	93%	3.86	100%	3.90	99%
1000	125	50	2.13	12%	2.39	35%	2.82	61%
1000	125	75	2.00	0%	2.02	2%	2.09	9%
1000	125	100	2.00	0%	2.00	0%	2.01	1%
1000	125	125	2.00	0%	2.00	0%	2.00	0%
2000	177	25	3.76	98%	3.93	99%	3.93	99%
2000	177	50	2.15	15%	2.56	47%	2.94	63%
2000	177	75	2.00	0%	2.02	2%	2.07	7%
2000	177	100	2.00	0%	2.00	0%	2.00	0%
2000	177	125	2.00	0%	2.00	0%	2.00	0%

Table 4.12: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 2, \phi = \frac{2}{3}, b = 0$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	3.07	77%	3.40	86%	3.51	92%
100	41	50	2.01	25%	2.15	35%	2.30	40%
100	41	75	1.72	27%	1.73	37%	1.90	32%
100	41	100	1.58	42%	1.58	40%	1.62	37%
100	41	125	1.51	47%	1.52	48%	1.44	56%
500	89	25	3.54	93%	3.68	95%	3.92	99%
500	89	50	2.20	20%	2.39	34%	2.80	68%
500	89	75	2.00	0%	2.02	2%	2.11	15%
500	89	100	2.00	0%	1.98	2%	1.97	3%
500	89	125	1.96	4%	1.96	4%	1.93	7%
1000	125	25	3.72	96%	3.83	98%	3.93	100%
1000	125	50	2.20	20%	2.54	45%	2.89	68%
1000	125	75	2.00	0%	2.02	2%	2.09	9%
1000	125	100	2.00	0%	2.00	0%	2.00	0%
1000	125	125	2.00	0%	2.00	0%	1.98	2%
2000	177	25	3.82	98%	3.91	99%	4.00	100%
2000	177	50	2.30	28%	2.66	50%	3.10	76%
2000	177	75	2.00	0%	2.05	5%	2.16	16%
2000	177	100	2.00	0%	2.00	0%	2.01	1%
2000	177	125	2.00	0%	2.00	0%	2.00	0%

Table 4.13: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 2, \phi = 0, b = \frac{1}{3}$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	3.57	98%	3.83	99%	3.89	100%
100	41	50	2.50	48%	3.03	79%	3.17	89%
100	41	75	1.99	25%	2.45	48%	2.64	65%
100	41	100	1.75	25%	1.90	24%	2.33	48%
100	41	125	1.71	29%	1.80	30%	1.97	31%
500	89	25	3.86	98%	3.96	99%	3.98	100%
500	89	50	2.82	62%	3.41	87%	3.80	97%
500	89	75	2.14	14%	2.57	48%	3.02	74%
500	89	100	2.01	3%	2.11	11%	2.39	41%
500	89	125	2.00	0%	2.00	4%	2.12	12%
1000	125	25	3.98	100%	3.99	100%	4.00	100%
1000	125	50	2.81	63%	3.51	94%	3.85	98%
1000	125	75	2.15	15%	2.62	52%	3.18	79%
1000	125	100	2.00	0%	2.07	9%	2.45	39%
1000	125	125	2.00	0%	2.02	2%	2.07	6%
2000	177	25	3.99	100%	4.00	100%	4.00	100%
2000	177	50	2.98	69%	3.67	96%	3.92	99%
2000	177	75	2.13	12%	2.77	60%	3.24	82%
2000	177	100	2.03	3%	2.11	11%	2.37	31%
2000	177	125	2.00	0%	2.02	2%	2.05	5%

Table 4.14: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 2, \phi = 0, b = \frac{2}{3}$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	3.89	99%	3.96	100%	3.99	100%
100	41	50	3.43	94%	3.81	99%	3.98	100%
100	41	75	3.02	81%	3.51	95%	3.81	98%
100	41	100	2.61	61%	3.36	90%	3.63	98%
100	41	125	2.32	41%	2.94	82%	3.52	95%
500	89	25	4.00	100%	3.99	100%	4.00	100%
500	89	50	3.91	99%	3.99	100%	4.00	100%
500	89	75	3.48	96%	3.93	99%	3.98	99%
500	89	100	2.84	68%	3.74	96%	3.98	100%
500	89	125	2.40	36%	3.42	92%	3.86	97%
1000	125	25	4.00	100%	4.00	100%	4.00	100%
1000	125	50	3.95	100%	4.00	100%	4.00	100%
1000	125	75	3.69	97%	3.99	100%	4.00	100%
1000	125	100	2.89	68%	3.81	99%	3.97	100%
1000	125	125	2.30	30%	3.32	82%	3.93	100%
2000	177	25	4.00	100%	4.00	100%	4.00	100%
2000	177	50	4.00	100%	4.00	100%	4.00	100%
2000	177	75	3.82	98%	3.99	100%	4.00	100%
2000	177	100	2.91	71%	3.84	98%	4.00	100%
2000	177	125	2.35	35%	3.36	89%	3.87	98%

Table 4.15: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 2, \phi = \frac{1}{3}, b = \frac{1}{3}$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	3.61	94%	3.76	95%	3.89	99%
100	41	50	2.52	52%	2.95	75%	3.24	87%
100	41	75	1.99	21%	2.24	44%	2.68	56%
100	41	100	1.76	26%	2.04	22%	2.32	35%
100	41	125	1.77	25%	1.76	33%	2.00	34%
500	89	25	3.90	98%	3.94	98%	3.99	100%
500	89	50	2.77	69%	3.40	87%	3.81	100%
500	89	75	2.12	12%	2.65	59%	3.10	79%
500	89	100	2.00	2%	2.10	12%	2.53	48%
500	89	125	1.98	2%	2.03	5%	2.13	12%
1000	125	25	3.95	99%	3.97	99%	4.00	100%
1000	125	50	2.95	72%	3.53	92%	3.92	99%
1000	125	75	2.14	14%	2.55	45%	3.24	83%
1000	125	100	2.03	3%	2.12	14%	2.37	35%
1000	125	125	2.00	0%	2.00	0%	2.15	15%
2000	177	25	3.98	100%	4.00	100%	4.00	100%
2000	177	50	2.91	63%	3.74	96%	3.92	97%
2000	177	75	2.21	19%	2.70	58%	3.46	92%
2000	177	100	2.00	0%	2.14	14%	2.58	47%
2000	177	125	2.00	0%	2.03	3%	2.17	17%

Table 4.16: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 1; \phi, b = 0; a = \frac{1}{3}$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	2.92	98%	3.12	99%	3.41	99%
100	41	50	1.44	42%	1.84	71%	2.11	83%
100	41	75	1.15	15%	1.28	26%	1.47	42%
100	41	100	1.01	1%	1.05	5%	1.14	14%
100	41	125	0.98	2%	1.01	1%	1.05	7%
500	89	25	3.35	98%	3.73	100%	3.83	100%
500	89	50	1.68	55%	2.07	81%	2.45	88%
500	89	75	1.09	9%	1.20	20%	1.57	51%
500	89	100	1.01	1%	1.06	6%	1.11	10%
500	89	125	1.00	0%	1.00	0%	1.01	1%
1000	125	25	3.53	99%	3.79	100%	3.96	100%
1000	125	50	1.70	59%	2.20	81%	2.73	93%
1000	125	75	1.08	8%	1.24	23%	1.57	47%
1000	125	100	1.00	0%	1.03	3%	1.09	11%
1000	125	125	1.00	0%	1.00	0%	1.02	2%
2000	177	25	3.82	100%	3.98	100%	4.00	100%
2000	177	50	1.86	66%	2.53	89%	2.96	93%
2000	177	75	1.13	13%	1.41	36%	1.63	49%
2000	177	100	1.01	1%	1.01	1%	1.17	16%
2000	177	125	1.00	0%	1.01	1%	1.00	0%

Table 4.17: Average $\hat{k}$ (Av.$\hat{k}$); and Rate of Error (RoE)
$k = 1; \phi, b = 0; a = \frac{2}{3}$

T	$\widetilde{T}$	N	Low-Noise		Mod.-Noise		High-Noise	
			Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE	Av.$\hat{k}$	RoE
100	41	25	3.04	95%	3.08	96%	3.39	100%
100	41	50	1.53	48%	1.91	72%	2.11	85%
100	41	75	1.12	12%	1.25	25%	1.57	54%
100	41	100	1.02	2%	1.11	11%	1.21	20%
100	41	125	1.00	0%	1.00	0%	1.07	7%
500	89	25	3.53	98%	3.67	99%	3.86	100%
500	89	50	1.63	54%	2.09	81%	2.53	87%
500	89	75	1.08	8%	1.25	23%	1.46	45%
500	89	100	1.00	0%	1.03	3%	1.17	17%
500	89	125	1.00	0%	1.00	0%	1.03	3%
1000	125	25	3.59	99%	3.85	100%	3.94	100%
1000	125	50	1.71	58%	2.23	78%	2.68	90%
1000	125	75	1.15	15%	1.32	27%	1.67	58%
1000	125	100	1.00	0%	1.09	9%	1.18	18%
1000	125	125	1.00	0%	1.01	1%	1.06	6%
2000	177	25	3.77	100%	3.97	100%	3.98	100%
2000	177	50	2.00	75%	2.57	84%	3.16	97%
2000	177	75	1.13	13%	1.35	32%	1.86	67%
2000	177	100	1.02	2%	1.13	13%	1.23	20%
2000	177	125	1.00	0%	1.02	2%	1.00	0%

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227. 21, 23, 154, 176, 191, 192, 203, 221

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23):1806–1813. 20, 21, 23, 154, 192, 208, 221

Altavilla, C., Giacomini, R., and Ragusa, G. (2017). Anchoring the yield curve using survey expectations. *Journal of Applied Econometrics*, 32(6):1055–1068. 84

Altissimo, F. and Corradi, V. (2003). Strong rules for detecting the number of breaks in a time series. *Journal of Econometrics*, 117:207–244. 160

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. 4

Aminikhanghahi, S. and Cook, D. J. (2017). A Survey of Methods for Time Series Change Point Detection. *Knowledge and Information Systems*, 51(2):339–367. 205

Anderson, B. and Deistler, M. (2008). Generalized linear dynamic factor models - a structure theory. In *2008 47th IEEE Conference on Decision and Control*, pages 1980–1985. 11

Andreou, E. and Ghysels, E. (2006). Monitoring disruptions in financial markets. *Journal of Econometrics*, 135(1):77–124. 160

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817–858. 65

Aue, A., Hörmann, S., Horváth, L., Hušková, M., and Steinebach, J. G. (2012). Sequential testing for the stability of high-frequency portfolio betas. *Econometric Theory*, 28(4):804–837. 161

Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046 – 4087. 170

Aue, A. and Horváth, L. (2004). Delay time in sequential detection of change. *Statistics & Probability Letters*, 67(3):221–231. 160

Aue, A., Horváth, L., Hušková, M., and Kokoszka, P. (2006). Change-point monitoring in linear models. *The Econometrics Journal*, 9(3):373–403. 161

Babu, G. J. and Singh, K. (1983). Inference on means using the bootstrap. *The Annals of Statistics*, 11:999–1003. 192

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171. 8, 25

Bai, J. and Li, K. (2012). Statistical Analysis of Factor Models of High Dimension. *The Annals of Statistics*, 40(1):436–465. 83

Bai, J. and Li, K. (2016). Maximum likelihood estimation and inference for approximate factor models of high dimension. *The Review of Economics and Statistics*, 98(2):298–309. 59, 83, 135

Bai, J. and Liao, Y. (2016). Efficient estimation of approximate factor models via penalized maximum likelihood. *Journal of Econometrics*, 191:1–18. 83

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221. 8, 20, 21, 23, 154, 173, 180, 181, 191, 208, 209, 221

Bai, J. and Ng, S. (2006). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica*, 74(4):1133–1150. 25

Bai, J. and Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, 25(1):52–60. 10, 25

Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3(2):89–163. 2

Bai, J. and Wang, P. (2015). Identification and Bayesian Estimation of Dynamic Factor Models. *Journal of Business & Economic Statistics*, 33(2):221–240. 35

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8(1):53–80. 2

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: from population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120. 49, 144, 148, 149, 150, 151

Baltagi, B. H., Kao, C., and Wang, F. (2017). Identification and estimation of a large factor model with structural instability. *Journal of Econometrics*, 197(1):87–100. 19, 153, 168, 169, 178

Bańbura, M., Giannone, D., and Lenza, M. (2015). Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections. *International Journal of Forecasting*, 31:739–756. 83

Bańbura, M. and Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, 29:133–160. 83

Banerjee, A., Marcellino, M., and Masten, I. (2005). Leading indicators for euro-area inflation and gdp growth. *Oxford Bulletin of Economics and Statistics*, 67:785–813. 4

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). *Now-Casting and the Real-Time Data Flow*, volume 2, chapter 4, pages 195–237. Elsevier. 4, 83

Barigozzi, M. (2020). *Dynamic Factor Models*. Lecture Notes, Università di Bologna. 2

Barigozzi, M., Cho, H., and Fryzlewicz, P. (2018a). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics*, 206(1):187–225. 19, 171

Barigozzi, M., Cho, H., and Owens, D. (2022). FNETS: Factor-adjusted network estimation and forecasting for high-dimensional time series. arXiv e-print. arXiv:2201.06110v2 [stat.ME] . 69

Barigozzi, M., Cuzzola, A., Grazzi, M., and Moschella, D. (2021a). Factoring in the micro: a transaction-level dynamic factor approach to the decomposition of export volatility. LEM Papers Series 2021/22, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy. 84

Barigozzi, M. and Hallin, M. (2015). Generalized dynamic factor models and volatilities: recovering the market volatility shocks. *The Econometrics Journal*, 19(1):C33–C60. 1, 3

Barigozzi, M., Hallin, M., and Soccorsi, S. (2018b). Identification of Global and Local Shocks in International Financial Markets via General Dynamic Factor Models. *Journal of Financial Econometrics*, 17(3):462–494. 30

Barigozzi, M., Lippi, M., and Luciani, M. (2021b). Large-dimensional Dynamic Factor Models: Estimation of Impulse-Response Functions with I(1) cointegrated factors. *Journal of Econometrics*, 221(2):455–482. 212, 213, 235

Barigozzi, M. and Luciani, M. (2021). Measuring the Output Gap using Large Datasets. *The Review of Economics and Statistics*, pages 1–45. 83

Barigozzi, M. and Luciani, M. (2022). Quasi Maximum Likelihood Estimation and Inference of Large Approximate Dynamic Factor Models via the EM algorithm.

arXiv e-print. arXiv:1910.03821v4 [math.ST]. 15, 16, 22, 25, 26, 59, 82, 83, 84, 85, 132, 133, 134, 135

Barigozzi, M. and Trapani, L. (2020). Sequential testing for structural stability in approximate factor models. *Stochastic Processes and their Applications*, 130(8):5149–5187. 20, 153, 156, 167, 168, 172, 185, 190, 191, 198, 199, 206

Barnard, G. A. (1959). Control charts and stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 21(2):239–271. 158

Barnett, W. A., Chauvet, M., and Leiva-Leon, D. (2016). Real-time nowcasting of nominal GDP with structural breaks. *Journal of Econometrics*, 191:312–324. 83

Bartholomew, D. J., Steele, F., Galbraith, J., and Moustaki, I. (2011). *Analysis of Multivariate Social Science Data*. Chapman and Hall/CRC, New York, 2nd edition. 6

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8:27–41. 193

Basseville, M. (1986). The two-models approach for the on-line detection of changes in ar processes. In Basseville, M. and Benveniste, A., editors, *Detection of Abrupt Changes in Signals and Dynamical Systems*, pages 169–215. Springer Berlin Heidelberg. 158

Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice Hall. 157

Bates, B. J., Plagborg-Møller, M., Stock, J. H., and Watson, M. W. (2013). Consistent factor estimation in dynamic factor models with structural instability. *Journal of Econometrics*, 177(2):289–304. Dynamic Econometric Modeling and Forecasting. 17, 18

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213. Data Mining for Software Trustworthiness. 209

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83. 209

Berkes, I., Gombay, E., Horváth, L., and Kokoszka, P. (2004). Sequential Change-point Detection in GARCH(p,q) Models. *Econometric Theory*, 20:1140–1167. 160

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120(1):387–422. 232

Bifet, A. and Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26–28, 2007, Minneapolis, MN*, pages 443–448. 207, 208

Bifet, A., Gavaldà, R., Holmes, G., and Pfahringer, B. (2017). *Machine Learning for Data Streams: with Practical Examples in MOA*. Adaptive Computation and Machine Learning . The MIT Press, Cambridge, MA. 205

Bracewell, R. N. (2000). *The Fourier transform and its applications*. McGraw Hill, Boston. 115

Breitung, J. and Eickmeier, S. (2006). Dynamic factor models. *AStA Advances in Statistical Analysis*, 90(1):27–42. 2

Breitung, J. and Eickmeier, S. (2011). Testing for structural breaks in dynamic factor models. *Journal of Econometrics*, 163(1):71–84. 18, 19, 153

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. SIAM. 10, 14, 25, 229

Brillinger, D. R. (1993). The digital rainbow: Some history and applications of numerical spectrum analysis. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 21(1):1–19. 22

Brodsky, B. (2012). Sequential Change-Point Detection in State-Space Models. *Sequential Analysis*, 31(2):145–171. 161

Burns, A. F. and Mitchell, W. C. (1946). *Measuring Business Cycles*. National Bureau of Economic Research. 3

Carlstein, E. (1986). The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *The Annals of Statistics*, 14:1171–1179. 193, 194, 210

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276. 21, 154

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica*, 51(5):1281–1304. 8

Chan, H. (2017). Optimal sequential detection in multi-stream data. *The Annals of Statistics*, 45(6):2736–2763. 164, 166, 206

Chen, L., Dolado, J. J., and Gonzalo, J. (2014). Detecting big structural breaks in large factor models. *Journal of Econometrics*, 180(1):30–48. 19, 153

Chen, Y., Wang, T., and Samworth, R. J. (2021). Inference in high-dimensional online changepoint detection. arXiv e-print. arXiv:2111.01640v1 [stat.ME]. 164

Chen, Y., Wang, T., and Samworth, R. J. (2022). High-dimensional, multiscale online changepoint detection. *Journal of the Royal Statistical Society Series B*, 84(1):234–266. 162, 164, 165, 166, 167, 209

Cheng, X., Liao, Z., and Schorfheide, F. (2016). Shrinkage estimation of high-dimensional factor models with structural instabilities. *Review of Economic Studies*, 83(4):1511–1543. 19, 153

Cho, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics*, 10:2000–2038. 171

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 77(2):475–507. 170, 171

Chu, C.-S. J., Hornik, K., and Kuan, C.-M. (1995). MOSUM Tests for Parameter Constancy. *Biometrika*, 82(3):603–617. 167

Chu, C.-S. J., Stinchcombe, M., and White, H. (1996). Monitoring Structural Change. *Econometrica*, 64(5):1045–1065. 152, 155, 159, 160, 172, 205, 208

Connor, G. and Korajczyk, R. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, 15(3):373–394. 8

Coroneo, L., Giannone, D., and Modugno, M. (2016). Unspanned macroeconomic factors in the yield curve. *Journal of Business & Economic Statistics*, 34(3):472–485. 84

Corradi, V. and Swanson, N. R. (2014). Testing for structural stability of factor augmented forecasting models. *Journal of Econometrics*, 182(1):100–118. 19, 153, 169, 178, 193

Deistler, M., Dunsmuir, W., and Hannan, E. J. (1978). Vector linear time series models: corrections and extensions. *Advances in Applied Probability*, 10(2):360–372. 59

Delle Chiaie, S., Ferrara, L., and Giannone, D. (2022). Common factors of commodity prices. *Journal of Applied Econometrics*, 37(3):461–476. 84

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. 16, 46, 48, 49, 144, 145, 147, 148

Dette, H., Eckle, T., and Vetter, M. (2020). Multiscale change point detection for dependent data. *The Annals of Statistics*, 47(4):1243–1274. 70

Dette, H. and Gösmann, J. (2018). Relevant change points in high dimensional time series. *Electronic Journal of Statistics*, 12(2):2578–2636. 69

Dette, H. and Gösmann, J. (2020). A likelihood ratio approach to sequential change point detection for a general class of parameters. *Journal of the American Statistical Association*, 115:1361–1377. 159

Dette, H., Pan, G., and Yang, Q. (2022). Estimating a change point in a sequence of very high-dimensional covariance matrices. *Journal of the American Statistical Association*, 117(537):444–454. 171

Diebold, F., Rudebusch, G., and Aruoba, S. B. (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of Econometrics*, 131(1-2):309–338. 3

Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods, and Applications with R Examples*. Texts in Statistical Science. Chapman and Hall/CRC, Boca Raton, FL, 1st edition. 147

Doz, C. and Fuleky, P. (2020). *Dynamic Factor Models*, chapter 2, pages 27–64. Springer International Publishing. 2, 3, 16, 23

Doz, C., Giannone, D., and Reichlin, L. (2006). A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models. Working Paper Series, No. 674. European Central Bank (ECB). 83

Doz, C., Giannone, D., and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1):188–205. Annals Issue on Forecasting. 12, 15, 16, 35, 46, 88, 122, 153

Doz, C., Giannone, D., and Reichlin, L. (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. *The Review of Economics and Statistics*, 94(4):1014–1024. 15, 16, 22, 25, 26, 82, 83

Dunsmuir, W. and Hannan, E. J. (1976). Vector linear time series models. *Advances in Applied Probability*, 8(2):339–364. 59

Dzhaparidze, K. (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. Springer Series in Statistics. Springer-Verlag, New York. 59

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26. 192

Eichenbaum, M. and Evans, C. L. (1995). Some empirical evidence on the effects of shocks to monetary policy on exchange rates. *The Quarterly Journal of Economics*, 110(4):975–1009. 4

Engle, R. and Watson, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, 76(376):774–781. 3

Engle, R. F. (1974). Band Spectrum Regression. *International Economic Review*, 15(1):1–11. 61

Evans, C. L., Liu, C. T., and Pham-Kanter, G. (2002). The 2001 recession and the Chicago Fed National Index: identifying business cycle turning points. *Economic Perspectives*, 26(Q III):26–43. 4

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. 4, 5, 8, 181

Favero, C. A., Marcellino, M., and Neglia, F. (2005). Principal components at work: the empirical analysis of monetary policy with large data sets. *Journal of Applied Econometrics*, 20(5):603–620. 25

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605. 157

Fiorentini, G., Galesi, A., and Sentana, E. (2018). A spectral EM algorithm for dynamic factor models. *Journal of Econometrics*, 205(1):249 – 279. 16, 22, 26, 27, 40, 50, 81, 82, 86

Forni, M. and Gambetti, L. (2010). The dynamic effects of monetary policy: A structural factor model approach. *Journal of Monetary Economics*, 57(2):203–216. 4, 231, 232, 233

Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2009). Opening the Black Box: Structural Factor Models with Large Cross Sections. *Econometric Theory*, 25(5):1319–1347. 10, 11, 12, 25, 68, 72, 122, 174, 180, 189, 214, 221, 232

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. 8, 9, 10, 14, 21, 22, 25, 30, 66, 73, 85, 88

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics*, 50(6):1243–1255. 4

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2004). The generalized dynamic factor model consistency and rates. *Journal of Econometrics*, 119(2):231–255. 8, 25

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*, 100:830–840. 8, 25, 85

Forni, M., Hallin, M., Lippi, M., and Zaffaroni, P. (2015). Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. *Journal of Econometrics*, 185(2):359–371. 10, 11, 85

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. 11, 15, 22, 25, 26, 60, 64, 65, 66, 69, 70, 71, 78, 85, 122, 138, 139, 140, 142

Forni, M. and Lippi, M. (2001). The Generalized Dynamic Factor Model: Representation Theory. *Econometric Theory*, 17(6):1113–1141. 8

Forni, M. and Lippi, M. (2011). The general dynamic factor model: One-sided representation results. *Journal of Econometrics*, 163(1):23–28. Factor Structures in Panel and Multivariate Time Series Data. 8, 11, 85

Fremdt, S. (2015). Page's sequential procedure for change-point detection in time series regression. *Statistics*, 49(1):128–155. 161

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42:2243–2281. 170

Geweke, J. (1977). The Dynamic Factor Analysis of Economic Time Series Models. *Latent Variables in Socioeconomic Models*, pages 365–383. 6

Geweke, J. F. and Singleton, K. J. (1981). Latent variable models for time series: A frequency domain approach with an application to the permanent income hypothesis. *Journal of Econometrics*, 17(3):287–304. 26

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676. 4, 83

Girshick, M. A. and Rubin, H. (1952). A Bayes Approach to a Quality Control Model. *The Annals of Mathematical Statistics*, 23(1):114 – 125. 157

Gómez, V. (2007). Wiener-Kolmogorov Filtering and Smoothing for Multivariate Series With State-Space Structure. *Journal of Time Series Analysis*, 28(3):361–385. 40

Gonçalves, S. and Perron, B. (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics*, 182(1):156–173. 193

Gonçalves, S. and Perron, B. (2020). Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics*, 218(2):476–495. 193

Gösmann, J., Kley, T., and Dette, H. (2021). A new approach for open-end sequential change point monitoring. *Journal of Time Series Analysis*, 42(1):63–84. 161

Gösmann, J., Stoehr, C., Heiny, J., and Dette, H. (2022). Sequential change point detection in high dimensional time series. *Electronic Journal of Statistics*, 16(1):3608 – 3671. 159, 167

Gourieroux, C. and Monfort, A. (1995). *Statistics and Econometrics Models*, volume 1 of *Themes in Modern Econometrics*. Cambridge University Press, Cambridge. 133

Grilli, V. and Roubini, N. (1996). Liquidity models in open economies: Theory and empirical evidence. *European Economic Review*, 40(3-5):847–859. 4

Groen, J. J. J., Kapetanios, G., and Price, S. (2013). Multivariate methods for monitoring structural change. *Journal of Applied Econometrics*, 28(2):250–274. 166

Gupta, M. R. and Chen, Y. (2011). Theory and use of the EM algorithm. *Foundations and Trends in Signal Processing*, 4(3):223–296. 148

Gustafsson, F. (2000). *Adaptive Filtering and Change Detection*. John Wiley & Sons, Ltd., Chichester, England. 205

Gut, A. and Steinebach, J. (2002). Truncated sequential change-point detection based on renewal counting processes. *Scandinavian Journal of Statistics*, 29(4):693–719. 160

Gut, A. and Steinebach, J. (2009). Truncated Sequential Change-point Detection based on Renewal Counting Processes II. *Journal of Statistical Planning and Inference*, 139(6):1921–1936. 160

Hall, P. (1985). Resampling a coverage pattern. *Stochastic Processes and their Applications*, 20:231–246. 193

Hall, P., Horowitz, J. L., and Jing, B. Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82:561–574. 194, 209

Hallin, M. and Lippi, M. (2013). Factor models in high-dimensional time series – time-domain approach. *Stochastic Processes and their Applications*, 123(7):2678–2695. A Special Issue on the Occasion of the 2013 International Year of Statistics. 8, 9, 10, 221

Hallin, M., Lippi, M., Barigozzi, M., Forni, M., and Zaffaroni, P. (2020). *Time Series in High Dimensions*. World Scientific. 2, 9, 14, 15

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. 20, 21, 23, 44, 69, 208, 221, 224, 232, 233, 234, 235

Han, X. and Inoue, A. (2014). Tests for Parameter Instability in Dynamic Factor Models. *Econometric Theory*, pages 1–36. 19, 153, 169, 178

Hannan, E. (1970). *Multiple time series*. John Wiley & Sons. 26, 40, 174, 189

Hannan, E. J. (1973). The asymptotic theory of linear time-series models. *Journal of Applied Probability*, 10(1):130–145. 59

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press. 6

He, Y., Kong, X. B., Trapani, L., and Yu, L. (2021). Online Change-point Detection for Matrix-valued Time Series with Latent Two-way Factor Structure. arXiv e-print. arXiv:2112.13479v1 [stat.ME] . 167, 168, 175

Hlávka, Z., Hušková, M., Kirch, C., and Meintanis, S. (2012). Monitoring changes in the error distribution of autoregressive models based on fourier methods. *TEST*, 21:605–634. 160

Hlávka, Z., Hušková, M., Kirch, C., and Meintanis, S. (2016). Bootstrap procedures for online monitoring of changes in autoregressive models. *Communications in Statistics Part B: Simulation and Computation*, 45(7):2471–2490. 160

Hörmann, S., Kidziński, Ł., and Hallin, M. (2015). Dynamic Functional Principal Component. arXiv e-print. arXiv:1210.7192v5 [math.ST]. 33

Horváth, L., Hušková, M., Kokoszka, P., and Steinebach, J. (2004). Monitoring changes in linear models. *Journal of Statistical Planning and Inference*, 126(1):225–251. 160

Horváth, L., Kokoszka, P., and Steinebach, J. (2007). On sequential detection of parameter changes in linear regression. *Statistics & Probability Letters*, 77(9):885–895. 160

Horváth, L., Kokoszka, P., and Wang, S. (2021). Monitoring for a change point in a sequence of distributions. *The Annals of Statistics*, 49(4):2271 – 2291. 162

Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis. *TEST*, 23:219 – 255. 156

Hušková, M. and Koubková (2005). Monitoring jump changes in linear models. *Journal of Statistical Research*, 39(2):51–70. 160

Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, 43:2451–2483. 69

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34:183–202. 6

Jungbacker, B. and Koopman, S. J. (2015). Likelihood-based dynamic factor analysis for measurement and forecasting. *The Econometrics Journal*, 18(2):C1–C21. 83

Jungbacker, B., Koopman, S. J., and van der Wel, M. (2011). Maximum likelihood estimation for dynamic factor models with missing data. *Journal of Economic Dynamics and Control*, 35(8):1358–1368. 83

Juvenal, L. and Petrella, I. (2015). Speculation in the Oil Market. *Journal of Applied Econometrics*, 30(4):621–649. 84

Kao, C., Trapani, L., and Urga, G. (2018). Testing for instability in covariance structures . *Bernoulli*, 24(1):740–771. 171

Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. In *Proceedings of the 30th VLDB Conference, Toronto, Canada*, pages 180–191. 205

Kirch, C. (2008). Bootstrapping sequential change-point tests. *Sequential Analysis*, 27(3):330–349. 160

Kirch, C. and Stoehr, C. (2022). Asymptotic delay times of sequential tests based on u-statistics for early and late change points. *Journal of Statistical Planning and Inference*, 221:114–135. 161

Kirch, C. and Tadjuidje Kamgaing, J. (2015). On the use of estimating functions in monitoring time series for change points. *Journal of Statistical Planning and Inference*, 161:25–49. 160

Krafty, R. T. and Collinge, W. O. (2013). Penalized multivariate Whittle likelihood for power spectrum estimation. *Biometrika*, 100(2):447–458. 50

Krämer, W., Ploberger, W., and Alt, R. (1988). Testing for structural change in dynamic models. *Econometrica*, 56:1355–1370. 159

Kulperger, R. and Yu, H. (2005). High moment partial sum processes of residuals in GARCH models and their applications. *The Annals of Statistics*, 33(5):2395–2422. 161

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241. 185, 193, 194, 195, 210

Lahiri, S. (2003). *Resampling Methods for Dependent Data*. Springer-Verlag, New York. 193, 194

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):613–644. 158, 164

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, 11(2):303–351. 158

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2):694–726. 12, 23, 176, 191, 230, 231

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918. 12

Lange, K. (2010). *Numerical analysis for statisticians*. Statistics and Computing. Springer, New York, 2nd edition. 144

Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229. 6

Leisch, F., Hornik, K., and Kuan, C.-M. (2000). Monitoring structural changes with the generalized fluctuation test. *Econometric Theory*, 16:835–854. 159

Linton, O. B., Tang, H., and Wu, J. (2022). A Structural Dynamic Factor Model for Daily Global Stock Market Returns. arXiv e-print. arXiv:2202.03638v4 [math.ST]. 84

Liu, R. Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of the Bootstrap*, pages 225–248. Wiley, New York. 193

Liu, W. and Wu, W. B. (2010). Asymptotics of Spectral Density Estimates. *Econometric Theory*, 26(4):1218–1245. 65, 69

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908. 156

Luciani, M. (2015). Monetary policy and the housing market: A structural factor analysis. *Journal of Applied Econometrics*, 30:199–218. 84

Luciani, M. (2020). Common and idiosyncratic inflation. Finance and Economics Discussion Series (FEDS) 2020-024. Federal Reserve, Washington. 83

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, Boston, 2nd edition. 144

Ma, S. and Su, L. (2018). Estimation of large dimensional factor models with an unknown number of breaks. *Journal of Econometrics*, 207(1):1–29. 19

Manner, H., Stark, F., and Wied, D. (2021). A monitoring procedure for detecting structural breaks in factor copula models. *Studies in Nonlinear Dynamics & Econometrics*, 25(4):171–192. 166, 167

Marcellino, M. and Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for german GDP. *Oxford Bulletin of Economics and Statistics*, 72:518–550. 83

Marcellino, M. and Sivec, V. (2016). Monetary, fiscal and oil shocks: Evidence based on mixed frequency structural FAVARs. *Journal of Econometrics*, 193(2):335–348. 83

McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley & Sons, Inc., New Jersey, 2 edition. 148

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, 31(6):471–482. 69

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, 97(2):419–433. 162, 163, 164, 166, 167

Merikoski, J. K. and Kumar, R. (2004). Inequalities for spreads of matrix sums and products. *Applied Mathematics E-Notes [electronic only]*, 4:150–159. 212

Modugno, M. (2013). Now-casting inflation using high frequency data. *International Journal of Forecasting*, 29:664–675. 83

Mody, A. and Taylor, M. P. (2007). Regional vulnerability: The case of east asia. *Journal of International Money and Finance*, 26(8):1292–1310. 3

Moustakides, G. V. (1986). Optimal Stopping Times for Detecting Changes in Distributions. *The Annals of Statistics*, 14(4):1379 – 1387. 156

Newey, W. K. and West, K. D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, 61(4):631–653. 65

Ng, C. T., Yau, C. Y., and Chan, N. H. (2015). Likelihood Inferences for High-Dimensional Factor Analysis of Time Series With Applications in Finance. *Journal of Computational and Graphical Statistics*, 24(3):866–884. 83

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479. 21, 23, 221, 235

Onatski, A. (2010). Determining the Number of Factors from Empirical Distribution of Eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016. 21, 154, 192

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115. 156, 157, 165

Parzen, E. (1957). On Consistent Estimates of the Spectrum of a Stationary Time Series. *Annals of Mathematical Statistics*, 28(2):329–348. 64

Pellegrino, F. (2022). Selecting time-series hyperparameters with the artificial jackknife. arXiv e-print. arXiv:2002.04697v4 [stat.ME] . 209

Ploberger, W., Krämer, W., and Kontrus, K. (1989). A new test for structural stability in the linear regression model. *Journal of Econometrics*, 40:307–318. 159

Politis, D. N. (2003). The Impact of Bootstrap Methods on Time Series Analysis. *Statistical Science*, 18(2):219–230. 193

Politis, D. N. and Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In *Exploring the Limits of the Bootstrap*, pages 263–270. Wiley, New York. 193

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313. 193

Pollak, M. (1985). Optimal detection of a change in distribution. *The Annals of Statistics*, 13:206–227. 157

Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *The Annals of Statistics*, 15:749–779. 157

Pollak, M. and Siegmund, D. (1985). A diffusion process and its application to detecting a change in the drift of a Brownian motion. *Biometrika*, 72:267–280. 157

Polunchenko, A. and Tartakovsky, A. (2012). State-of-the-Art in Sequential Change-Point Detection. *Methodology and Computing in Applied Probability*, 14:649–684. 156

Poor, H. and Hadjiliadis, O. (2008). *Quickest detection*. Cambridge University Press, United Kingdom. 156

Priestley, M. B. (1982). *Spectral Analysis and Time Series*, volume 1–2 of *Probability and Mathematical Statistics*. Academic Press, London, 1 edition. 62, 65, 79, 131, 142

Quah, D. and Sargent, T. (1993). *A Dynamic Index Model for Large Cross Sections*, pages 285–310. University of Chicago Press. 6, 26

Reis, R. and Watson, M. (2010). Relative Goods' Prices, Pure Inflation, and The Phillips Correlation. *American Economic Journal: Macroeconomics*, 2:128–157. 83

Ritov, Y. (1990). Decision Theoretic Optimality of the CUSUM Procedure. *The Annals of Statistics*, 18(3):1464 – 1469. 156

Sargent, T. and Sims, C. (1977). *Business Cycle Modeling Without Pretending to Have Too Much a Priori Economic Theory*, pages 45–109. Federal Reserve Bank of Minneapolis. 6, 26

Seber, G. A. F. (2008). *A Matrix Handbook for Statisticians*. John Wiley & Sons, Inc., New Jersey, 1st edition. 211

Shao, X. and Wu, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *The Annals of Statistics*, 35(4):1773–1801. 69

Shewhart, W. (1931). *Economic Control of Quality of Manufactured Product*. Bell Telephone Laboratories series. American Society for Quality Control. 156

Shiryaev, A. N. (1961). The Problem of the Most Rapid Detection of a Disturbance in a Stationary Process. *Soviet Mathematics – Doklady*, 2:795–799. 157

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, 8(1):22–46. 157

Shiryaev, A. N. (1965). Some exact formulas in a disorder process. *Theory of Probability and its Applications*, 10(3):348–354. 157

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review*, 36(5):975–1000. 4

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, 9:1187–1195. 192

Soh, Y. S. and Chandrasekaran, V. (2017). High-dimensional change-point estimation: Combining filtering with convex optimization. *Applied and Computational Harmonic Analysis*, 43(1):122–147. 206

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292. 2, 3, 5

Stock, J. and Watson, M. (1988). A probability model of the coincident economic indicators. *NBER Working Paper Series*. 7

Stock, J. and Watson, M. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293–335. 4, 232

Stock, J. and Watson, M. (2016). Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. volume 2 of *Handbook of Macroeconomics*, pages 415–525. Elsevier. 2, 23, 24

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179. 4, 8, 17, 18, 88, 173, 180, 221

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162. 4, 8, 221

Stock, J. H. and Watson, M. W. (2005). Implications of Dynamic Factor Models for VAR Analysis. Working Paper 11467. National Bureau of Economic Research. 10, 35

Sundberg, R. and Feldmann, U. (2016). Exploratory factor analysis - Parameter estimation and scores prediction with high-dimensional data. *Journal of Multivariate Analysis*, 148:49–59. 83

Tartakovsky, A., Nikiforov, I., and Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman and Hall/CRC, New York, 1st edition. 19, 156, 196

Tartakovsky, A. G., Rozovskii, B. L., Blažek, R. B., and Kim, H. (2006). Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, 3(3):252–293. 162, 163, 164, 166, 167

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622. 6

Trapani, L. (2015). Inference on the number of factors in factor models. In *21st International Panel Data Conference, CEU, Budapest*. 226, 229, 234

Trapani, L. (2018). A Randomized Sequential Procedure to Determine the Number of Factors. *Journal of the American Statistical Association*, 113(523):1341–1349. v, 21, 23, 174, 185, 220, 221, 222, 223, 228, 229, 231, 234, 235

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299. 206

Venkataraman, E. S. (1992). Consistency results in multiple change-point problems. Technical Report 24, Department of Statistics, Stanford University, Stanford. 170

Vostrikova, L. (1981). Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics – Doklady*, 24:55–59. 170

Wald, A. (1948). *Sequential Analysis*. Wiley & Sons, New York. 156, 157

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339. 156, 157

Wang, D., Yu, Y., and Rinaldo, A. (2017). Optimal covariance change point localization in high dimensions. arXiv e-print. arXiv:1712.09912v2 [math.ST]. 170

Wang, D., Yu, Y., and Rinaldo, A. (2021). Optimal covariance change point localization in high dimensions. *Bernoulli*, 27:554–575. 169, 171, 172

Watson, M. W. and Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23(3):385–400. 6

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25. 15

Wied, D. and Galeano, P. (2005). Monitoring correlation change in a sequence of random variables. *Journal of Statistical Planning and Inference*, 143:186–196. 159

Wiener, N. (1950). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. Technology Press of the Massachusetts Institute of Technology, Cambridge. 40

Willsky, A. and Jones, H. (1976). A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, 21(1):108–112. 158

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103. 46, 49, 132, 133, 134, 144, 145, 146, 147, 148, 149

Wu, W. (2011). Asymptotic theory for stationary processes. *Statistics and its Interface*, 4:207–226. 70

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154. 69

Wu, W. B. and Zaffaroni, P. (2018). Asymptotic Theory for Spectral Density Estimates of General Multivariate Time Series. *Econometric Theory*, 34(1):1–22. iv, 14, 15, 22, 26, 27, 64, 68, 69, 70, 138, 143

Xie, Y. and Siegmund, D. (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics*, 41(2):670–692. 164, 166, 206

Yamamoto, Y. (2016). Forecasting With Nonspurious Factors in U.S. Macroeconomic Time Series. *Journal of Business & Economic Statistics*, 34(1):81–106. 18

Yamamoto, Y. and Tanaka, S. (2015). Testing for factor loading structural change under common breaks. *Journal of Econometrics*, 189(1):187–206. 19, 153

Yu, Y. (2020). A review on minimax rates in change point detection and localisation. arXiv e-print. arXiv:2011.01857v1 [math.ST] . 172

Zangwill, W. I. (1969). *Nonlinear Programming: A Unified Approach*. Prentice-Hall, Englewood Cliffs, NJ. 144

Zeileis, A., Leisch, F., Kleiber, C., and Hornik, K. (2005). Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics*, 20:99–121. 159

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. 69

Zhou, M., Wang, H. J., and Tang, Y. (2015). Sequential change point detection in linear quantile regression models. *Statistics & Probability Letters*, 100:98–103. 160