Factor Modelling for High-Dimensional Time Series: A Dimension-Reduction Approach*

Clifford Lam and Qiwei Yao

Department of Statistics, The London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, U.K.

Abstract

Following a brief survey on the factor models for multiple time series in econometrics, we introduce a statistical approach from the viewpoint of dimension reduction. Our method can handle nonstationary factors. However under stationary settings, the inference is simple in the sense that both the number of factors and the factor loadings are estimated in terms of an eigenanalysis for a non-negative definite matrix, and is therefore applicable when the dimension of time series is in the order of a few thousands. Asymptotic properties of the proposed method are investigated under two settings: (i) the sample size goes to infinity while the dimension of time series is fixed; and (ii) both the sample size and the dimension of time series go to infinity together. In particular, our estimators for zero-eigenvalues enjoy faster convergence (or slower divergence) rates, hence making the estimation for the number of factors easier. In particular when the sample size and the dimension of time series go to infinity together, the estimators for the eigenvalues are no longer consistent. However our estimator for the number of the factor, which is based on the ratios of the estimated eigenvalues, still works fine. Furthermore, the estimation for both the number of factors and the factor loadings shows the so-called "blessing of dimensionality" property. A two-step procedure is investigated when the factors are of different degrees of strength. Numerical illustration with both simulated and real data is also reported.

Key words and phrases. Auto-covariance matrices, Bless of dimensionality, Curse of dimensionality, Eigenanalysis, Fast convergence rates, Idiosyncratic component, Multivariate time series, Nonstationarity, Ratio-based estimator, Strength of factors, White noise.

*Partially supported by an EPSRC research grant.

1 Introduction

The analysis of multivariate time series data is of increased interest and importance in the modern information age. Although the methods and the associate theory for univariate time series analysis are well developed and understood, the picture for the multivariate cases is less complete. Although the conventional univariate time series models (such as ARMA) and the associated time-domain and frequency-domain methods have been formally extended to multivariate cases, their usefulness is often limited. One may face serious issues such as the lack of model identification or flat likelihood functions. In fact vector ARMA models are seldom used directly in practice. Model regularization via, for example, reduced-rank structure, structural indices, scalar component models and canonical correlation analysis is most pertinent in modelling multivariate time series data. See Hannan (1970), Priestley (1981), Lütkepohl (1993), and Reinsel (1997).

In this paper we survey the recent developments in factor modelling for multivariate time series from a dimension-reduction viewpoint. By doing so, we have also developed some new results. Different from the factor analysis for independent observations, we search for the factors which drive the serial dependence of the original time series. Early attempts in this direction include Anderson (1963), Priestley et al.(1974), Brillinger (1981), Switzer and Green (1984), Peña and Box (1987), and Shapiro and Switzer (1989). More recent efforts focus on the inference when the dimension of time series is as large as or even greater than the sample size; see, for example, Chamberlain and Rothschild (1983), Bai (2003), Forni et al.(2000, 2004 and 2005). High-dimensional time series data are often encountered nowadays in many fields including finance, economics, environmental and medical studies. For example, understanding the dynamics of the returns of large number of assets is the key for asset pricing, portfolio allocation, and risk management. Panel time series are commonplace in studying economic and business phenomena. Environmental time series are often of a high dimension due to a large number of indices monitored across many different locations.

Factor models for high-dimensional time series have been featured noticeably in literature in econometrics and finance. In analyzing economic and financial phenomena, most econometric factor models seek to identify the *common factors* that affect the dynamics of most original component series. It is often meaningful in practice to separate these common factors from the so-called idiosyncratic noise components: each idiosyncratic noise component may at most affect the dynamics of a few original time series. An idiosyncratic noise series may well exhibit serial correlations, and may be a time series itself. This poses the difficulties in both model identification and inference. In fact the rigorous definition of the common factors and the idiosyncratic noise can only be established asymptotically when the dimension of time series tends to infinity; see Chamberlain and Rothschild (1983). Hence those econometric factor models are only asymptotically identifiable. See also section 2 below.

Our approach is more statistical and is from a dimension-reduction point of view. Our model

is similar to that of Peña and Box (1987). Different from the aforementioned econometric factor models, we decompose a high-dimensional time series into two parts: a dynamic part driven by, hopefully, a lower-dimensional factor time series, and a static part which is a vector white noise. Since the white noise exhibits no serial correlations, the decomposition is unique in the sense that both the number of factors (i.e. the dimension of the factor process) and the factor loading space in our model are identifiable. Such a conceptually simple decomposition also makes the statistical inference easier. Our setting allows the factor process to be non-stationary; see Pan and Yao (2008) and also section 4.1 below. However, under the stationary condition, the inference is simple in the sense that the estimation for both the number of factors and the loadings is carried out in an eigenanalysis for a non-negative definite matrix, and is therefore applicable when the dimension of time series is in the order of a few thousands. Asymptotic properties of the proposed method are investigated under two settings: (i) the sample size goes to infinity while the dimension of time series is fixed; and (ii) both the sample size and the dimension of time series go to infinity together. In particular, our estimators for zero-eigenvalues enjoy the faster convergence (or slower divergence) rates, from which the proposed ratio-based estimator for the number of factors benefits immensely. In fact, the estimation for the number of factors shows the so-called "blessing of dimensionality" property at its clearest.

The rest of the paper is organized as follows. Section 2 gives a brief survey on the econometric factor models, including the generalized dynamic factor models. Section 3 presents our statistical factor models. Although the two approaches are based on different viewpoints, they do reconcile with each other in practical data analysis (see section 3.2). The estimation methods for the statistical models are introduced in section 4, including a brief account for the cases with non-stationary factors. Their asymptotic theory (for the stationary cases only) are presented in section 5. Section 6 deals with the cases when different factors are of different strength, for which a two-step estimation procedure is preferred. Simulation results are inserted whenever appropriate to illustrate the various properties of the proposed methods. The analysis of two large real data sets is reported in section 7. Most mathematical proofs are relegated to the Appendix. Throughout the paper, new theoretical results are presented as theorems and corollaries. Some relevant results from other papers are presented as propositions.

2 Econometrics models: a brief introduction

2.1 Common factors and idiosyncratic components

Time series factor models have been constantly featured in econometrics literature. They are used to model different economic and financial phenomena, including, among others, asset pricing (Ross 1976) and allocation (Pesaran and Zaffaroni 2008), yield curves (Chib and Ergashev 2009), macroeconomic behaviour such as sector-effect or regional effect from disaggregated data (Quah

and Sargent 1993, Forni and Reichlin 1998), macroeconomic forecasting (Stock and Watson 1998, 2002), capital accumulation and growth (Chudik and Pesaran 2009) and consumer theory (Bai 2003).

Among different factor models in econometric literature, one predominate feature is to represent a $p \times 1$ time series $\mathbf{y}_t$ as the sum of two unobservable parts:

$$\mathbf{y}_t = \mathbf{f}_t + \boldsymbol{\xi}_t, \tag{2.1}$$

where $\mathbf{f}_t$ is a factor term driven by r common factors with r smaller or much smaller than p, and $\boldsymbol{\xi}_t$ is an idiosyncratic term which consists of p idiosyncratic components. For example, in the context of modelling comovements of economic activities, the comovements such as the oscillation between upturn phases and depression phases are presented by the common factor term $\mathbf{f}_t$, while the idiosyncratic dynamic behaviour for each component series is contained in $\boldsymbol{\xi}_t$. Obviously it is extremely appealing and also important to isolate the common factors from many idiosyncratic components in practice. Since $\boldsymbol{\xi}_t$ is not necessarily a white noise, both the identification and the inference for decomposition (2.1) is inevitably challenging. In fact $\mathbf{f}_t$ and $\boldsymbol{\xi}_t$ are only asymptotically identifiable when p, i.e. the number of components $\mathbf{y}_t$, tends to ∞; see Chamberlain and Rothschild (1983) and also section 2.2 below.

2.2 Generalized dynamic factor models

The dynamic-factor model was proposed by Sargent & Sims (1977) and Geweke (1977). It assumes that in the decomposition (2.1) each component of $\mathbf{f}_t$ is a sum of r uncorrelated moving average processes driven, respectively, by r common factors (see (2.2) below). Furthermore it requires that $\mathbf{f}_t$ and $\boldsymbol{\xi}_t$ are uncorrelated with each other, and all the idiosyncratic components (i.e. the components of $\boldsymbol{\xi}_t$) are also uncorrelated. Chamberlain and Rothschild (1983) and Chamberlain (1983) proposed an approximate static factor model in which the factor term is of the form $\mathbf{f}_t = \mathbf{A}\mathbf{x}_t$, where $\mathbf{x}_t$ is an $r \times 1$ factor process. Since no lagged values of $\mathbf{x}_t$ is involved explicitly, $\mathbf{x}_t$ is coined as a *static* factor. The model is *approximate* in the sense that the idiosyncratic components of $\boldsymbol{\xi}_t$ may be correlated with each other now, which makes the model more practical. Chamberlain and Rothschild (1983) proved that the Ross' asset pricing theorem still holds under this approximate factor model, i.e. if there exists no arbitrage opportunities, the risk premium of an asset is determined by its factor loadings in a particularly simple manner (i.e. an asymptotic linear function).

Let y_{jt} and ξ_{jt} be the j-th component of $\mathbf{y}_t$ and $\boldsymbol{\xi}_t$ respectively. Combining the above two approaches, Forni *et al.* (2000) proposed a generalized dynamic factor model which may be expressed as follows:

$$y_{jt} = b_{j1}(L)u_{1t} + \cdots + b_{jr}(L)u_{rt} + \xi_{jt}, \quad t = 0, \pm 1, \pm 2, \cdots, \ j = 1, \cdots, p, \tag{2.2}$$

where $u_{1t}, \cdots, u_{rt}$ are r uncorrelated white noise and are called common (dynamic) factors, $b_{j1}(\cdot), \cdots, b_{jr}(\cdot)$ are polynomial functions, L denotes the backshift operator, and the idiosyncratic components ξ_{jt} are stationary in t and are uncorrelated with the common factors. Note that for each $1 \leq i \leq r$, $b_{ji}(L)u_{it}$ is a moving average process driven by white noise u_{it}.

In (2.2) only y_{jt} is observable. To make the terms on the right hand side of (2.2) identifiable, Forni *et al.* (2000) introduced the following asymptotic condition when p, the number of components of $\mathbf{y}_t$, tends to infinity. We write $\mathbf{y}_{tp} \equiv \mathbf{y}_t$ and $\boldsymbol{\xi}_{tp} \equiv \boldsymbol{\xi}_t$ to highlight the length p of both vectors $\mathbf{y}_t$ and $\boldsymbol{\xi}_t$.

Assumption. *As $p \to \infty$, it holds almost surely on $[-\pi, \pi]$ that all the eigenvalues of the spectral density matrix of $\boldsymbol{\xi}_{tp}$ are uniformly bounded, and only the r largest eigenvalues of the spectral density matrix of $(\mathbf{y}_{tp} - \boldsymbol{\xi}_{tp})$ converge to ∞.*

The intuition behind the above assumption is clear: As $p \to \infty$, the number of components of $\mathbf{y}_{tp}$ whose dynamics depends on all the r common factors also tends to infinity, while each of the idiosyncratic components only affects at most a finite number of components of $\mathbf{y}_{tp}$. The idea to use this asymptotic argument to identify the model was initiated by Chamberlain and Rothschild (1983), although they dealt with the covariance matrices of $\mathbf{y}_{tp}$ and $\boldsymbol{\xi}_{tp}$ instead of their spectral density matrices. The latter reflects the strength of the autocorrelations across all time lags. Under the above assumption, Forni *et al.* (2000) showed that model (2.2) is asymptotically identifiable; see Proposition 1 below.

Proposition 1 *As $p \to \infty$, it holds almost surely on $[-\pi, \pi]$ that all the r largest eigenvalues of the spectral density matrix of $\mathbf{y}_{tp}$ converge to ∞, and the $(r+1)$-th largest eigenvalue is uniformly bounded.*

The statistical inference for model (2.2) is not a trivial matter. By assuming r as given, the estimation for the common factors (as well as the idiosyncratic components) is resolved by applying the dynamic principal component analysis (Brillinger 1981). Namely eigenanalysis is performed on an estimated spectral density matrix $\boldsymbol{\Sigma}_n(\theta)$ of $\mathbf{y}_{tp}$, which is defined for $\theta \in [-\pi, \pi]$. Note that $\boldsymbol{\Sigma}_n(\theta)$ is typically obtained by applying smoothing to the periodogram matrices of the observations $\mathbf{y}_{1p}, \cdots, \mathbf{y}_{np}$, and the subscript n here indicates that it is an estimator based on n observations. A 'projection' of $\mathbf{y}_{tp}$ onto a dynamic space spanned by the r eigenvectors of $\boldsymbol{\Sigma}_n(\theta)$ corresponding to the r largest eigenvalues is taken as an estimator for $\mathbf{y}_{tp} - \boldsymbol{\xi}_{tp}$. Note both of those eigenvectors and eigenvalues are functions of the frequency $\theta \in [-\pi, \pi]$, and the projection was defined as a mean square limit of an appropriate Fourier expansion. Then each component of the projection is a sum of r uncorrelated moving average processes. For further details of this estimation procedure, we refer to Forni *et al.* (2000). See also Forni *et al.* (2004, 2005) and Deistler *et al.* (2009).

A more challenging task is to determine the number of the common factors r. Proposition 1 indicates clearly that r is only asymptotically identifiable when $p \to \infty$. In practice p is always a fixed finite number. Note that r is the number of eigenvalues which diverge in contrast to the other $p - r$ eigenvalues which remain bounded as $p \to \infty$. 'There is no way a slowly diverging sequence (divergence, under the model, can be arbitrarily slow) can be told from an eventually bounded sequence (for which the bound can be arbitrarily large)', as pointed out in Forni et al (2000, pp.547). Therefore the estimation for r often relies on some empirical approaches. Hallin and Liska (2007) proposed an information criterion to determine r. See also Bai and Ng (2002, 2007), adopting a similar approach in dealing with a factor model of different specifications.

3 Statistical models

3.1 Models

If we are interested in the linear dynamic structure of $\mathbf{y}_t$ only, conceptually we may think that $\mathbf{y}_t$ consists of two parts: a dynamic component driven by, hopefully, a low-dimensional process and a static part (i.e. a white noise). This leads to the decomposition:

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{3.1}$$

where $\mathbf{x}_t$ is an $r \times 1$ latent process with (unknown) $r \leq p$, $\mathbf{A}$ is a $p \times r$ unknown constant matrix, and $\boldsymbol{\varepsilon}_t \sim \text{WN}(\boldsymbol{\mu}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)$ is a vector white noise process. When r is much smaller than p, we achieve an effective dimension-reduction, as then the serial dependence of $\mathbf{y}_t$ is driven by that of a much lower-dimensional process $\mathbf{x}_t$. We call $\mathbf{x}_t$ a factor process. The setting (3.1) may be traced back at least to Peña and Box (1987); see also its further development in dealing with cointegrated factors in Peña and Poncela (2006).

Since none of the elements on the RHS of (3.1) are observable, we have to characterize them further to make them identifiable. First we assume that no linear combinations of $\mathbf{x}_t$ are white noise, as any such component can be absorbed into $\boldsymbol{\varepsilon}_t$. We also assume that the rank of $\mathbf{A}$ is r. Otherwise (3.1) may be expressed equivalently in terms of a lower-dimensional factor. Furthermore, since (3.1) is unchanged if we replace $(\mathbf{A}, \mathbf{x}_t)$ by $(\mathbf{AH}, \mathbf{H}^{-1}\mathbf{x}_t)$ for any invertible $r \times r$ matrix $\mathbf{H}$, we may assume that the columns of $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$ are orthonormal, i.e., $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$, where $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. Note that even with this constraint, $\mathbf{A}$ and $\mathbf{x}_t$ are not uniquely determined in (3.1), as the aforementioned replacement is still applicable for any orthogonal $\mathbf{H}$. However the factor loading space, i.e. the r-dimensional linear space spanned by the columns of $\mathbf{A}$, denoted by $\mathcal{M}(\mathbf{A})$, is uniquely defined.

We summarize into condition C1 all the assumptions introduced so far.

C1. In model (3.1), $\boldsymbol{\varepsilon}_t \sim \text{WN}(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$, $\mathbf{c}'\mathbf{x}_t$ is not white noise for any constant $\mathbf{c} \in \mathcal{R}^p$. Furthermore $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$.

The key in the statistical inference for model (3.1) is to estimate $\mathbf{A}$, or more precisely $\mathcal{M}(\mathbf{A})$. Once we have obtained an estimator, say, $\widehat{\mathbf{A}}$, a natural estimator for the factor process is

$$\widehat{\mathbf{x}}_t = \widehat{\mathbf{A}}'\mathbf{y}_t, \tag{3.2}$$

and the resulting residuals are

$$\widehat{\boldsymbol{\varepsilon}}_t = (\mathbf{I}_d - \widehat{\mathbf{A}}\widehat{\mathbf{A}}')\mathbf{y}_t. \tag{3.3}$$

The dynamic modelling for $\mathbf{y}_t$ is achieved via $\widehat{\mathbf{y}}_t = \widehat{\mathbf{A}}\widehat{\mathbf{x}}_t$, and such a modelling for $\widehat{\mathbf{x}}_t$. A parsimonious fitting for $\widehat{\mathbf{x}}_t$ may be obtained by rotating $\widehat{\mathbf{x}}_t$ appropriately (Tiao and Tsay 1989). Such a rotation is equivalent to replacing $\widehat{\mathbf{A}}$ by $\widehat{\mathbf{A}}\mathbf{H}$ for an appropriate $r \times r$ orthogonal matrix $\mathbf{H}$. Note that $\mathcal{M}(\widehat{\mathbf{A}}) = \mathcal{M}(\widehat{\mathbf{A}}\mathbf{H})$, and the residuals (3.3) are unchanged with such a replacement.

3.2 Reconciling to econometric models

One of the attractive features of the econometric factor models is the separation of the idiosyncratic components from the 'common factors'; see (2.1). In principle the factor term $\mathbf{A}\mathbf{x}_t$ in statistical decomposition (3.1) contains both the common factor term $\mathbf{f}_t$ and those components of $\boldsymbol{\xi}_t$ which exhibit serial dependence in (2.1). Although the 'common factors' driving $\mathbf{f}_t$ can only be identified asymptotically when $p \to \infty$, in practice we may identify those components of $\mathbf{x}_t$ such that each of them drives the serial dependence of the majority components of $\mathbf{y}_t$, or alternatively, those components of $\mathbf{x}_t$ such that each of them only affects a few components of $\mathbf{y}_t$. In fact this is all we may achieve in practice with a fixed p.

Put $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$ and $\mathbf{x}_t = (x_{1t}, \cdots, x_{rt})'$. Then model (3.1) can be written as

$$\mathbf{y}_t = \mathbf{a}_1 x_{1t} + \cdots + \mathbf{a}_r x_{rt} + \boldsymbol{\varepsilon}_t. \tag{3.4}$$

Hence the number of non-zero coefficients of $\mathbf{a}_j$ is the number of components of $\mathbf{y}_t$ which are affected by the factor x_{jt}. Obviously such a characterization is only meaningful if $\mathbf{x}_t$ is uniquely defined. To this end, we replace $\mathbf{x}_t$ by its principal components. This effectively replace $(\mathbf{A}, \mathbf{x}_t)$ in (3.4) by $(\mathbf{A}\boldsymbol{\Gamma}, \boldsymbol{\Gamma}'\mathbf{x}_t)$, where $\boldsymbol{\Gamma}$ is an $r \times r$ orthogonal matrix, and its r columns are the eigenvectors of $\mathrm{Var}(\mathbf{x}_t)$. Note that after such an replacement, $\mathrm{Cov}(x_{it}, x_{jt}) = 0$ for any $i \neq j$. If the eigenvalues of $\mathrm{Var}(\mathbf{x}_t)$ are all different, Lemma 1 below indicates that such an $\mathbf{x}_t$ is unique subject to the permutations and the reflections (i.e. replace x_{jt} by $-x_{jt}$) of its components.

Lemma 1 *Let $\mathbf{A}_1\mathbf{z}_1 = \mathbf{A}_2\mathbf{z}_2$, where, for $i = 1, 2$, $\mathbf{A}_i$ is $p \times r$ matrix, $\mathbf{A}_i'\mathbf{A}_i = \mathbf{I}_r$, and $\mathbf{z}_i = (z_{i1}, \cdots, z_{ir})'$ is $r \times 1$ random vector with uncorrelated components, and $Var(z_{i1}) > \cdots > Var(Z_{ir})$. Furthermore $\mathcal{M}(\mathbf{A}_1) = \mathcal{M}(\mathbf{A}_2)$. Then $z_{1j} = \pm z_{2j}$ for $j = 1, \cdots, r$.*

Proof. Let $\boldsymbol{\Sigma}_i = \mathrm{Var}(\mathbf{z}_i)$. Then $\mathbf{A}_1\boldsymbol{\Sigma}_1\mathbf{A}_1' = \mathbf{A}_2\boldsymbol{\Sigma}_2\mathbf{A}_2'$. Hence

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_2\boldsymbol{\Gamma}', \tag{3.5}$$

where $\mathbf{\Gamma} = \mathbf{A}_1'\mathbf{A}_2$ is an orthogonal matrix. This is due to fact that $\mathcal{M}(\mathbf{A}_1) = \mathcal{M}(\mathbf{A}_2)$, which implies that $\mathbf{A}_2 = \mathbf{A}_1\mathbf{\Gamma}$. Since $\mathbf{\Sigma}_2$ is diagonal, (3.5) implies that the diagonal elements of $\mathbf{\Sigma}_2$ are the eigenvalues of matrix $\mathbf{\Sigma}_1$. As $\mathbf{\Sigma}_1$ itself is also diagonal, and the diagonal elements in both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are arranged in descending order, it must hold that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Thus $\mathbf{\Gamma} = \mathbf{A}_1'\mathbf{A}_2$ is also a diagonal matrix with the elements on the main diagonal being 1 or -1. Then the required conclusion follows from the equality $\mathbf{z}_1 = \mathbf{A}_1'\mathbf{A}_2\mathbf{z}_2 = \mathbf{\Gamma}\mathbf{z}_2$. ■

4 Estimation for A and r

Below we outline two methods for estimating $\mathbf{A}$ (and also r). The method via expanding the white noise space is more general. It can handle nonstationary factors. Unfortunately it involves solving some nonlinear optimization problems with up to $(p-1)$ variables, therefore is only applicable with a moderately large p. If we are prepared to entertain the stationarity condition (see C2 in section 4.2 below), the problem boils down to finding eigenvalues and eigenvectors for a $p \times p$ non-negative definite matrix. Furthermore r is the number of non-zero eigenvalues of this matrix. Hence the method can be applied to the cases with p in the order of a few thousands.

4.1 Estimation with nonstationary factors

Our goal is to estimate $\mathcal{M}(\mathbf{A})$, or, equivalently, its orthogonal complement $\mathcal{M}(\mathbf{B})$, where $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_{p-r})$ is a $p \times (p-r)$ matrix for which $(\mathbf{A}, \mathbf{B})$ forms a $p \times p$ orthogonal matrix, i.e. $\mathbf{B}'\mathbf{A} = 0$ and $\mathbf{B}'\mathbf{B} = \mathbf{I}_{p-r}$ (see also C1). We call $\mathcal{M}(\mathbf{B})$ the white noise space.

It follows from (3.1) that

$$\mathbf{B}'\mathbf{y}_t = \mathbf{B}'\boldsymbol{\varepsilon}_t, \tag{4.1}$$

implying that for any $1 \le j \le p-r$, $\{\mathbf{b}_j'\mathbf{y}_t,\ t = 0, \pm 1, \cdots\}$ is a white noise process. Hence, we may search for mutually orthogonal directions $\mathbf{b}_1, \mathbf{b}_2, \cdots$ one by one such that the projection of $\mathbf{y}_t$ on each of those directions is a white noise. We stop the search when such a direction is no longer available, and take $p-k$ as the estimated value of r, where k is the number of directions obtained in the search. Below we outline a schematic algorithm to search for those $\mathbf{b}_j$ and r. For further details on the implementation of this scheme, we refer to Pan and Yao (2008). Note that once we have obtained an estimator $\widehat{\mathbf{B}}$, the columns of $\widehat{\mathbf{A}}$ may be taken as any r orthonormal eigenvectors of matrix $(\mathbf{I}_p - \widehat{\mathbf{B}}\widehat{\mathbf{B}}')$ corresponding to the eigenvalue 1.

White noise expansion algorithm:

Step 1. Find unit vector $\widehat{\mathbf{b}}_1 \in \mathcal{R}^p$ such that $\{\widehat{\mathbf{b}}_1'\mathbf{y}_t\}$ is most likely to be white noise among all unit vectors in $\mathcal{R}^p$. Test the null hypothesis that $\{\widehat{\mathbf{b}}_1'\mathbf{y}_t\}$ is a white noise. If it cannot be rejected, proceed to Step 2. Otherwise set $\widehat{r} = 0$ and terminate the algorithm.

Step 2. For $k = 1, \cdots, p$, search for unit vector $\widehat{\mathbf{b}}_{k+1} \in \mathcal{R}^p$ such that $\{\widehat{\mathbf{b}}'_{k+1}\mathbf{y}_t\}$ is most likely to be white noise among all directions in $\mathcal{R}^p$ perpendicular to $\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_k$. Test the null hypothesis that $\{\widehat{\mathbf{b}}'_{k+1}\mathbf{y}_t\}$ is a white noise. If the null hypothesis is rejected, terminate the algorithm with $\widehat{r} = p - k$ and $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_k)$.

In principle, this method works under condition C1 which is rather weak. For example, we do not require condition C2 in section 4.2 below, as we only make use of the property that any projection of $\mathbf{y}_t$ in $\mathcal{M}(\mathbf{B}) = \mathcal{M}(\mathbf{A})^\perp$ is white noise. The theoretical exploration of the method in Pan and Yao (2008) requires more regularity conditions. Intuitively some of those conditions are not essential.

4.2 Estimation under stationarity

A much simpler method is available with the stationarity condition on the factor process as follows.

C2. $\mathbf{x}_t$ is weakly stationary, and $\mathrm{Cov}(\mathbf{x}_t, \boldsymbol{\varepsilon}_{t+k}) = 0$ for any $k \geq 0$.

In most factor modelling literature, $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_s$ are assumed to be uncorrelated for any t and s. Condition C2 requires only that the future white noise components are uncorrelated with the factors up to the present. This enlarges the model capacity substantially. Put

$$\boldsymbol{\Sigma}_y(k) = \mathrm{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t), \qquad \boldsymbol{\Sigma}_x(k) = \mathrm{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t), \qquad \boldsymbol{\Sigma}_{x\varepsilon}(k) = \mathrm{Cov}(\mathbf{x}_{t+k}, \boldsymbol{\varepsilon}_t).$$

It follows from (3.1) and C2 that

$$\boldsymbol{\Sigma}_y(k) = \mathbf{A}\boldsymbol{\Sigma}_x(k)\mathbf{A}' + \mathbf{A}\boldsymbol{\Sigma}_{x\varepsilon}(k), \qquad k \geq 1. \tag{4.2}$$

For a prescribed integer $k_0 \geq 1$, define

$$\mathbf{M} = \sum_{k=1}^{k_0} \boldsymbol{\Sigma}_y(k)\boldsymbol{\Sigma}_y(k)'. \tag{4.3}$$

Then $\mathbf{M}$ is a $p \times p$ non-negative matrix. It follows from (4.2) that $\mathbf{MB} = 0$, i.e. the columns of $\mathbf{B}$ are the eigenvectors of $\mathbf{M}$ corresponding to zero-eigenvalues. Hence conditions C1 and C2 imply:

The factor loading space $\mathcal{M}(\mathbf{A})$ are spanned by the eigenvectors of $\mathbf{M}$ corresponding to its non-zero eigenvalues, and the number of the non-zero eigenvalues is r.

We take the sum in the definition of $\mathbf{M}$ to accumulate the information from different time-lags. This is useful especially when the sample size n is small. We use the non-negative definite matrix $\boldsymbol{\Sigma}_y(k)\boldsymbol{\Sigma}_y(k)'$ (instead of $\boldsymbol{\Sigma}_y(k)$) to avoid the cancellation of the information from different lags. This is guaranteed by the fact that for any matrix $\mathbf{C}$, $\mathbf{MC} = 0$ if and only if $\boldsymbol{\Sigma}_y(k)'\mathbf{C} = 0$ for all $1 \leq k \leq k_0$. We tend to use small k_0, as the autocorrelation is often at its strongest at the small time lags. On the other hand, adding more terms will not alter the value of r, although the

estimation for $\mathbf{\Sigma}_y(k)$ with large k is less accurate. The simulation results reported in Lam, Yao and Bathia (2011) also confirms that the estimation for $\mathbf{A}$ and r, defined below, is not sensitive to the choice of k_0.

To estimate $\mathcal{M}(\mathbf{A})$, we only need to perform the eigenanalysis on

$$\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)', \tag{4.4}$$

where $\widehat{\mathbf{\Sigma}}_y(k)$ denotes the sample covariance matrix of $\mathbf{y}_t$ at lag k. Then the estimator $\widehat{r}$ for the number of factors is taken as the number of non-zero eigenvalues $\widehat{\mathbf{M}}$, and the columns of the estimated factor loading matrix $\widehat{\mathbf{A}}$ are the $\widehat{r}$ orthonormal eigenvectors of $\widehat{\mathbf{M}}$ corresponding to its $\widehat{r}$ largest eigenvalues.

Due to the random fluctuation in a finite sample, the estimates for the zero-eigenvalues of $\mathbf{M}$ may not be 0 exactly. A common practice is to plot all the estimated eigenvalues in an descending order, and look for a cut-off value $\widehat{r}$ such that the $(\widehat{r}+1)$-th largest eigenvalue is substantially smaller than the $\widehat{r}$ largest eigenvalues. This is effectively an eyeball-test. The ratio-based estimator defined below may be viewed as an enhanced eyeball-test, based on the same idea as Wang (2010). In fact this ratio-based estimator benefits from the faster convergence rates of the estimators for the zero-eigenvalues; see Proposition 2 in section 5.1 below, and also Theorems 1 and 2 in section 5.2 below. The other available methods for determining r includes the information criteria approaches of Bai and Ng (2002, 2007), and Hallin and Liska (2007), the bootstrap approach of Bathia, Yao Ziegelmann (2010).

A ratio-based estimator for r. We define an estimator for the number of factors r as follows:

$$\widehat{r} = \arg\min_{1\le i\le R} \widehat{\lambda}_{i+1}/\widehat{\lambda}_i, \tag{4.5}$$

where $r < R < p$ is a constant.

In practice we may use, for example, $R = p/2$. We cannot extend the search up to p, as the minimum eigenvalue of $\widehat{M}$ is likely to be practically 0, especially when n is small and p is large. It is worthy noting that when p and n are in the same order, the estimators for eigenvalues are no longer consistent. However the ratio-based estimator (4.5) still works well. See Theorem 2(iii) below.

The above estimation methods for $\mathbf{A}$ and r can be extended to those nonstationary time series for which a generalized lag-k autocovariance matrix is well defined (see, e.g. Peña and Poncela (2006)). In fact, the methods are still applicable when the weak limit of the generalized lag-k autocovariance matrix

$$\widehat{\mathbf{S}}_y(k) = n^{-\alpha}\sum_{t=1}^{n-1}(\mathbf{y}_{t+k} - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})'$$

exists for $1 \le k \le k_0$, where $\alpha > 1$ is a constant. Further developments on those lines will be reported elsewhere. For the factor modelling for high-dimensional volatility processes based on a

similar idea, we refer to Pan et al.(2011) and Tao et al.(2011).

5 Theoretical results

Conventional asymptotic properties are established under the setting that the sample size n tends to ∞ and everything else remains fixed. Modern time series analysis encounters the situation when the number of time series p is as large as, or ever larger than, the sample size n, e.g. panel data in economics and ecology, and asset prices in financial market. Then the asymptotic properties established under the setting when both n and p tend to ∞ are more relevant. We analyze these two different settings in sections 5.1 and 5.2 respectively.

5.1 Asymptotics when $n \to \infty$ and p fixed

We first consider the asymptotic properties under the assumption that $n \to \infty$ and p is fixed. These properties reflect the behaviour of our estimation method in the cases when n is large and p is small. We introduce some regularity conditions first. Let $\lambda_1, \cdots, \lambda_p$ be the eigenvalues of matrix $\mathbf{M}$.

C3. $\mathbf{y}_t$ is strictly stationary and ψ-mixing with the mixing coefficients $\psi(\cdot)$ satisfying the condition that $\sum_{t\geq 1} t\psi(t)^{1/2} < \infty$. Furthermore $E\{|\mathbf{y}_t|^4\} < \infty$ elementwisely.

C4. $\lambda_1 > \cdots > \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_p$.

Section 2.6 of Fan and Yao (2003) gives a compact survey on the mixing properties of time series. The use of the ψ-mixing condition in C3 is for technical convenience. Note that $\mathbf{M}$ is a non-negative definite matrix. All its eigenvalues are non-negative. C4 assumes that its r non-zero eigenvalues are distinct from each other. This condition is not essential. However it substantially simplifies the presentation on the convergence of the estimated eigenvalues and eigenvectors in Proposition 2 below. For example, under C4 the unit eigenvector of $\mathbf{M}$ corresponding to the eigenvalue λ_j, for $j = 1, \cdots, r$, is uniquely determined (up to a factor -1). Let $\boldsymbol{\gamma}_j$ be such an eigenvector. We denote $(\widehat{\lambda}_1, \widehat{\boldsymbol{\gamma}}_1), \cdots, (\widehat{\lambda}_p, \widehat{\boldsymbol{\gamma}}_p)$ the p pairs of eigenvalue and eigenvector of matrix $\widehat{\mathbf{M}}$: the eigenvalues $\widehat{\lambda}_j$ are arranged in the descending order, and the eigenvectors $\widehat{\boldsymbol{\gamma}}_j$ are orthonormal. Furthermore it may go without explicit statement that $\widehat{\lambda}_j$ may be replaced by $-\widehat{\lambda}_j$ in order to match the direction of λ_j for $1 \leq j \leq r$.

Proposition 2 *Let conditions C1-C4 hold. Then as $n \to \infty$ (but p fixed), it holds that*

(i) $|\widehat{\lambda}_j - \lambda_j| = O_P(n^{-1/2})$ *and* $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\| = O_P(n^{-1/2})$ *for* $j = 1, \cdots, r$, *and*

(ii) $\widehat{\lambda}_j = O_P(n^{-1})$ *for* $j = r+1, \cdots, p$.

The proof of the above proposition is in principle similar to that of Theorem 1 in Bathia, Yao and Ziegelmann (2010), is therefore omitted. If we let $\mathbf{A} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_r)$ and $\widehat{\mathbf{A}} = (\widehat{\boldsymbol{\gamma}}_1, \cdots, \widehat{\boldsymbol{\gamma}}_r)$, assuming r is known, it follows from Proposition 2(i) that $\widehat{\mathbf{A}} - \mathbf{A} = O_P(n^{-1/2})$ elementwisely.

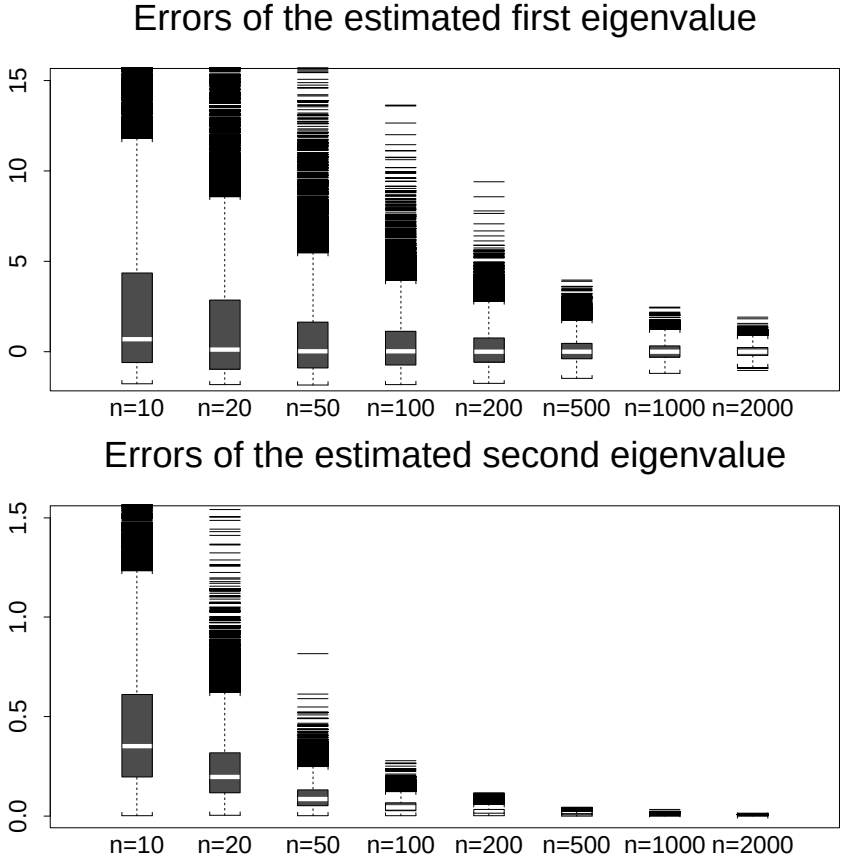

Figure 1: *Boxplots for the errors in estimating the first and the second eigenvalues of* $\mathbf{M}$ *with* $p = 4, r = 1$.

The fast convergence rate of the estimators for the zero-eigenvalues is non-standard (Proposition 2(ii)). To further circumstance this property, we conduct a simulation: we set in model (3.1) $p = 4$, $r = 1$, $\mathbf{A}' = (1, 0, 0, 0)$, $\boldsymbol{\varepsilon}_t$ are independent $N(0, \mathbf{I}_4)$, and $\mathbf{x}_t = x_t$ is an AR(1) process defined by

$$x_{t+1} = 0.7x_t + e_t, \tag{5.1}$$

where e_t are independent $N(0, 1)$. Let $\mathbf{M}$ be defined as in (4.3) with $k_0 = 1$. Then $\mathbf{M}$ is a 4×4 matrix with one non-zero eigenvalue $\lambda_1 = 1.884$ and three zero eigenvalues. We let sample size vary from $n = 20$ to $n = 2000$. We draw 10,000 samples from such a model with each fixed sample size. For each sample, we compute the eigenvalues of $\widehat{M}$, denote them as $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \widehat{\lambda}_3 \geq \widehat{\lambda}_4$. Fig. 1 displays the boxplots of the errors $\widehat{\lambda}_1 - \lambda_1$ and $\widehat{\lambda}_2$ over 10,000 samples with different n. Though $\widehat{\lambda}_2$ is in fact the maximum of the three estimated values (for the three zero-eigenvalues), they are much smaller than the errors $\widehat{\lambda}_1 - \lambda_1$. Fig. 2 plots the normalized histograms of $\sqrt{n}(\widehat{\lambda}_1 - \lambda_1)$ together with their kernel density estimators. As n increases, the distribution of $\sqrt{n}(\widehat{\lambda}_1 - \lambda_1)$ stabilizes. In fact, the distributions with $n \geq 500$ look alike with each other. However, the normalized factor $\sqrt{n}$ is too small to stabilize the distribution of $\widehat{\lambda}_2 - \lambda_2 = \widehat{\lambda}_2$. Fig. 3 shows that

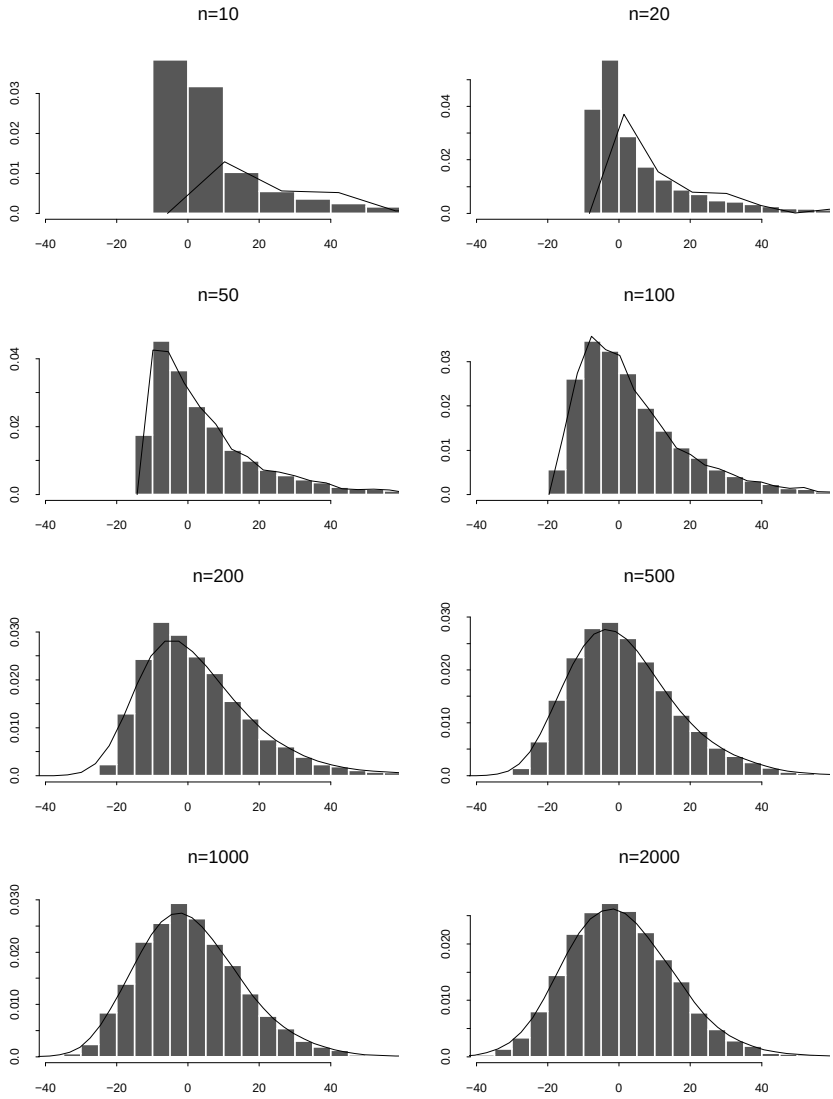

Figure 2: *Standardized histograms and kernel density estimators for* $\sqrt{n}(\widehat{\lambda}_1 - \lambda_1)$, $\widehat{\lambda}_1$ *is the largest eigenvalue of* $\widehat{\mathbf{M}}$, *and* $p = 4, r = 1$.

the distribution of $n\widehat{\lambda}_2$ stabilizes as soon as $n \geq 50$. This indicates that the estimators for zero-eigenvalues not only exhibit the fast convergence rate at n, they may attain the limit distribution much earlier than the estimators for the non-zero eigenvalues when n increases.

Below we give an intuitive explanation on the results in Proposition 2. Note that we may count the number of the zero-eigenvalues of $\mathbf{\Sigma}_y(k)$ in order to determine the number of factors r, as $\mathbf{B}\mathbf{\Sigma}_y(k) = 0$. But we look into $\mathbf{M}$ instead in order to accumulate the information over different time lags, and further we define $\mathbf{M}$ as a quadratic function of $\mathbf{\Sigma}_y(k)$ to avoid the cancellation of the information at different lags. For example when $k_0 = 1$, an eigenvalue of $\mathbf{M}$ is the square of an eigenvalue of $\mathbf{\Sigma}_y(1)$. In this sense, we estimate an eigenvalue θ of $\mathbf{\Sigma}_y(k)$ in the form of $g(\theta) = \theta^2$

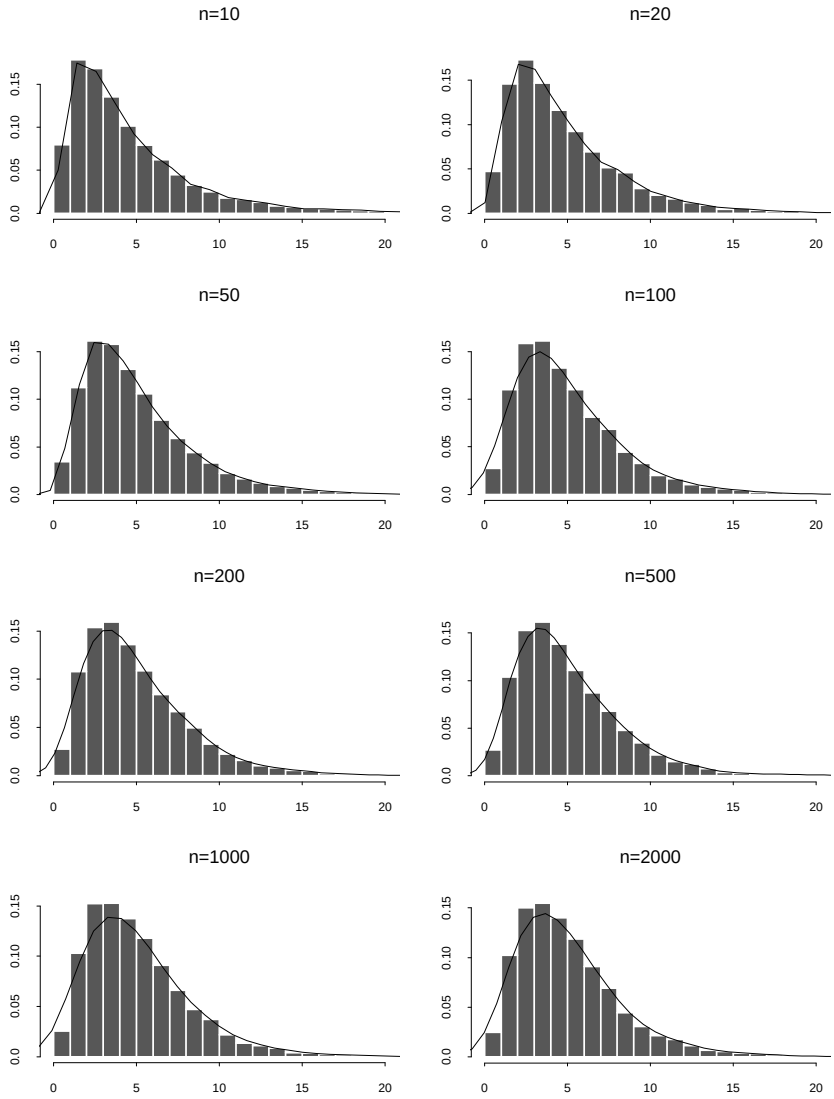

Figure 3: *Standardized histograms and kernel density estimators for* $n\widehat{\lambda}_2$, $\widehat{\lambda}_2$ *is the second largest eigenvalue of* $\widehat{\mathbf{M}}$, *and* $p = 4, r = 1$.

via $\mathbf{M}$. Suppose we have $\widehat{\theta} - \theta = O_P(n^{-1/2})$, and

$$g(\widehat{\theta}) - g(\theta) = \dot{g}(\theta)(\widehat{\theta} - \theta) + \frac{1}{2}\ddot{g}(\theta)(\widehat{\theta} - \theta)^2\{1 + o_P(1)\},$$

Hence $g(\widehat{\theta}) - g(\theta) = O_P(n^{-1/2})$ provided $\dot{g}(\theta) \neq 0$. However when $\theta = 0$, $\dot{g}(\theta) = 0$. Hence $\widehat{\theta}^2 = g(\widehat{\theta}) - g(\theta) = O_P\{(\widehat{\theta} - \theta)^2\} = O_P(n^{-1})$.

5.2 Asymptotics when $n \to \infty$, $p \to \infty$ and r fixed

5.2.1 An introductory example

To highlight the radically different behaviour when p diverges together with n, we first repeat the simulation in section 5.1 above but now with $\mathbf{A}' = (1, \cdots, 1)$, the sample size $n = 50, 100, 200, 400,$

800, 1600 and 3200, and the dimension fixed at the half sample size, i.e. $p = n/2$. For each setting, we draw 200 samples. The boxplots of the errors $\widehat{\lambda}_i - \lambda_i$, $i = 1, \cdots, 6$ are depicted in Fig. 4. Note that $\lambda_i = 0$ for $i \geq 2$, since $r = 1$. The figure shows that those estimation errors do not converge to 0. In fact those errors seem to increase when n (and also $p = n/2$) increases. Therefore the classic asymptotic theory (i.e. $n \to \infty$ and p fixed) such as Proposition 1 in section 5.1 above is irrelevant when p increases together with n. In spite of the lack of consistency in estimating the eigenvalues, the ratio-based estimator for the number of factors $r(= 1)$ defined in (4.5) works perfectly fine for this example, as shown in Figure 5. In fact it is always the case that $\widehat{r} \equiv 1$ in all our replicated simulation even when the sample size is as small as $n = 50$; see Fig.5.

5.2.2 Properties of the estimated factor loading matrix

To develop the relevant asymptotic theory, we introduce some notation first. For any matrix $\mathbf{G}$, let $\|\mathbf{G}\|$ be the square root of the maximum eigenvalue of $\mathbf{G}\mathbf{G}'$, and $\|\mathbf{G}\|_{\min}$ be the square root of the smallest positive eigenvalue of $\mathbf{G}\mathbf{G}'$. We write $a \asymp b$ if $a = O(b)$ and $b = O(a)$. Recall $\mathbf{\Sigma}_x(k) = \mathrm{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t)$ and $\mathbf{\Sigma}_{x\epsilon}(k) = \mathrm{Cov}(\mathbf{x}_{t+k}, \boldsymbol{\varepsilon}_t)$. Some regularity conditions are now in order.

C5. For a constant $\delta \in [0, 1]$, it holds that $\|\mathbf{\Sigma}_x(k)\| \asymp p^{1-\delta} \asymp \|\mathbf{\Sigma}_x(k)\|_{\min}$.

C6. For $k = 0, 1, \cdots, k_0$, $\|\mathbf{\Sigma}_{x\epsilon}(k)\| = o(p^{1-\delta})$.

Remark 1. (i) Condition C5 looks unnatural. It is derived from a set of natural conditions coupled with the standardization $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$. Since $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$ is $p \times r$ and $p \to \infty$ now, it is natural to let the norm of each column of $\mathbf{A}$, before standardizing to $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$, tend to ∞ as well. To this end, we assume that

$$\|\mathbf{a}_j\|^2 \asymp p^{1-\delta_j}, \quad j = 1, \cdots, r, \tag{5.2}$$

where $\delta_j \in [0, 1]$ are constants. We take δ_j as a measure of the strength of the factor x_{tj}; see also (3.4). We call x_{tj} a strong factor when $\delta_j = 0$, and a weak factor when $\delta_j > 0$. Since r is fixed, it is also reasonable to assume that for $k = 0, 1, \cdots, k_0$,

$$|\mathbf{\Sigma}_x(k)| \neq 0, \tag{5.3}$$

Then condition C5 is entailed by the standardization $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$ under conditions (5.3) and (5.2) with $\delta_j = \delta$ for all j; see Lam, Yao and Bathia (2010), and also the discussion on Proposition 3 below.

(ii) The condition assumed on $\mathbf{\Sigma}_{x\epsilon}(k)$ in C6 requires that the correlation between $\mathbf{x}_{t+k}$ ($k \geq 0$) and $\boldsymbol{\varepsilon}_t$ is not too strong. In fact under a natural condition that $\mathbf{\Sigma}_{x\epsilon}(k) = O(1)$ elementwisely, it is implied by (5.2) and the standardization $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$ that $\|\mathbf{\Sigma}_{x,\epsilon}(k)\| = O(p^{1-\delta/2})$. A more general rate of convergence in the presence of stronger correlation between $\{\mathbf{x}_t\}$ and $\{\boldsymbol{\varepsilon}_t\}$ has been established in Lam, Yao and Bathia (2010). For the sake of simplicity, we only present the results under condition C6 in this paper.

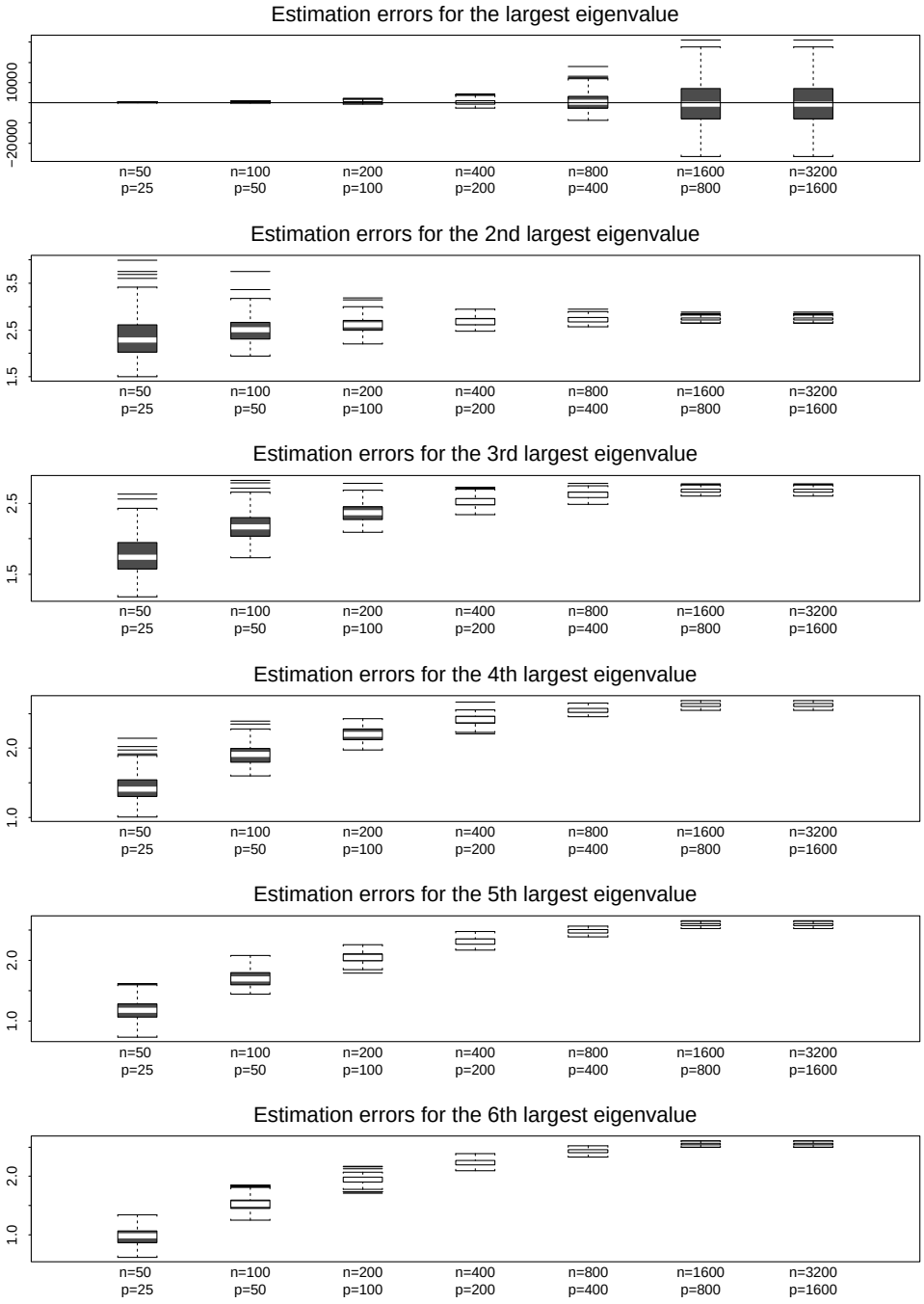

Figure 4: *Boxplots for the errors in estimating the first six eigenvalues of* $\mathbf{M}$ *with* $r = 1$ *and all the factor loading coefficients being 1.*

Proposition 3 *Let conditions C1-C6 hold and* $h_n \equiv p^{\delta} n^{-1/2} \to 0$. *Then as* $n \to \infty$ *and* $p \to \infty$, *it holds that*

$$\|\widehat{\mathbf{A}} - \mathbf{A}\| = O_P(h_n) \xrightarrow{P} 0,$$

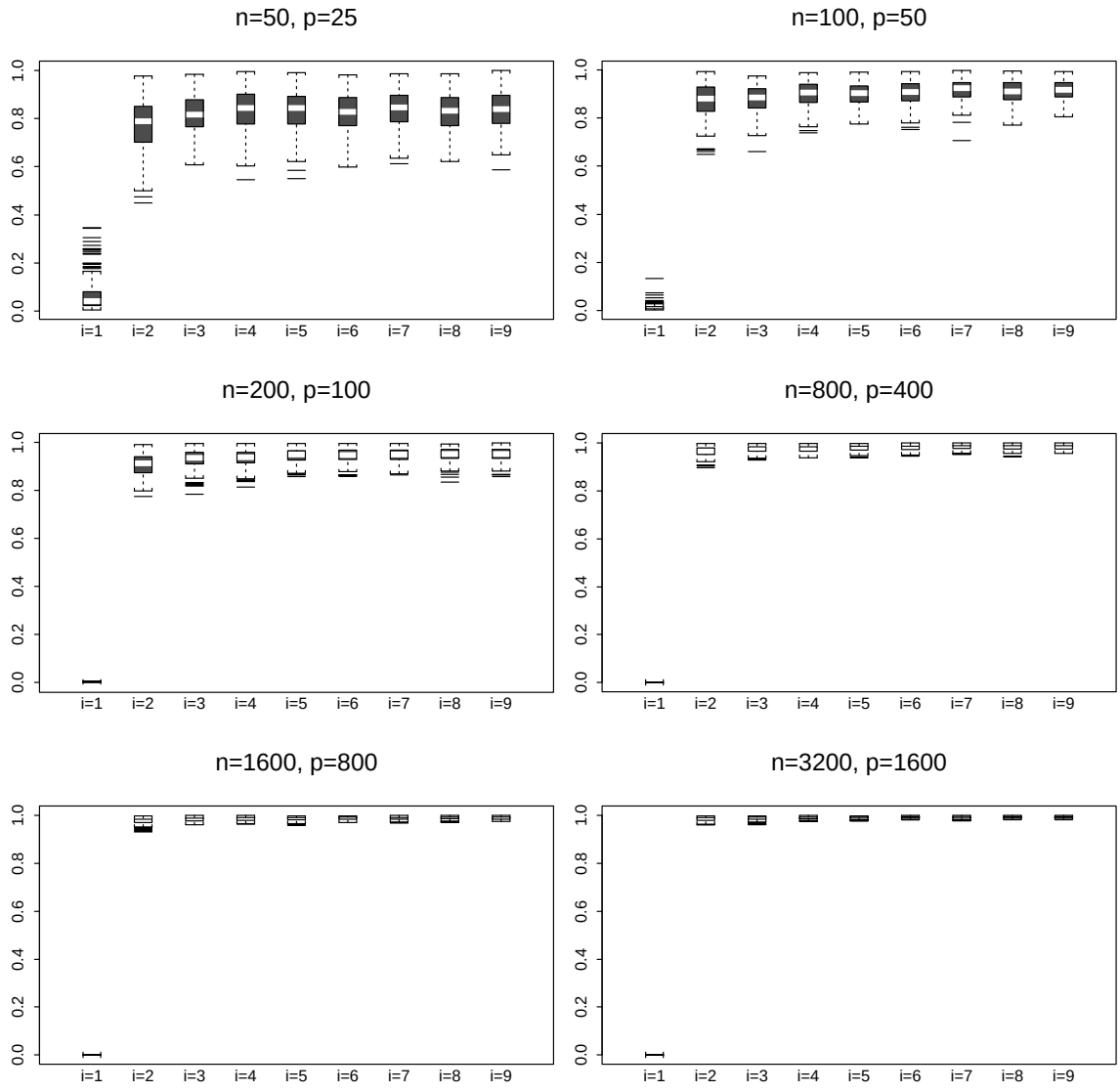

Figure 5: *Boxplots for the ratios* $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$*, with* $r = 1$ *and all the factor loading coefficients being 1.*

When all the factors are strong (i.e. $\delta = 0$), $\|\widehat{\mathbf{A}} - \mathbf{A}\| = O_P(n^{-1/2})$, i.e. the convergence rate is independent of p. Note that $\mathbf{A}$ is a $p \times r$ matrix. The number of estimated parameters increases together with the dimension p. This is one of the examples where the 'curse of dimensionality' is canceled out by the 'blessing of the dimensionality'. Since the r factors are fixed, we gain more information on $\mathbf{x}_t$ from more component series of $\mathbf{y}_t$ when p increases. The condition $\delta = 0$ ensures that all the components of $\mathbf{x}_t$ are indeed *common* factors in the sense that each of them affect the majority components of $\mathbf{y}_t$. Hence when p increases, the added components of $\mathbf{y}_t$ indeed bring in more information on $\mathbf{x}_t$.

The proof of Proposition 3 can be found in Lam, Yao and Bathia (2010). It also reports a simulation study which shows that the accuracy in estimating $\mathbf{A}$ (with finite samples) is indeed

independent of p when all factors are strong. The 'blessing of dimensionality' is also observed in estimating the precision matrix of $\mathbf{y}_t$, i.e. the inverse of the covariance matrix of $\mathbf{y}_t$. However, it does not apply to the estimation for the covariance matrix itself. All those asymptotic results will be reported elsewhere.

5.2.3 Properties of the estimated eigenvalues

Now we deal with the convergence rates of the estimated eigenvalues, and establish the results in the same spirit as Proposition 2. Of course the convergence (or divergence) rate for each estimator $\widehat{\lambda}_i$ is slower, as the number of estimated parameters goes to infinity now.

Theorem 1 *Let conditions C1-C6 hold and $h_n = p^{\delta} n^{-1/2} \to 0$. Then as $n \to \infty$ and $p \to \infty$, it holds that*

(i) $|\widehat{\lambda}_i - \lambda_i| = O_P(p^{2-\delta} n^{-1/2})$ *for* $i = 1, \cdots, r$, *and*

(ii) $\widehat{\lambda}_j = O_P(p^2 n^{-1})$ *for* $j = r+1, \cdots, p$.

The corollary below follows immediately from the above theorem and the fact that $\lambda_j \asymp p^{2-2\delta}$ for $j = 1, \cdots, r$ (see condition C5 and Remark 1(i)).

Corollary 1 *Under the condition of Theorem 1, it holds that*

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \asymp 1 \ \text{ for } j = 1, \cdots, r-1, \quad \text{and } \ \widehat{\lambda}_{r+1}/\widehat{\lambda}_r = O_P(h_n^2) \xrightarrow{P} 0.$$

The proof of Theorem 1 is relegated to the Appendix. Obviously when p is fixed, Theorem 1 formally reduces to Proposition 2. Some remarks are now in order.

Remark 2. (i) Corollary 1 implies that the plot of ratios $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$, $i = 1, 2, \cdots$, will drop sharply at $i = r$. This provides a useful theoretical underpinning for the estimator $\widehat{r}$ defined in (4.5). Unfortunately we are unable to derive an explicit asymptotic expression for the ratios $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ with $i > r$, although we make the following conjecture:

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \xrightarrow{P} 1, \qquad j = (k_0+1)r+1, \cdots, (k_0+1)r+K, \tag{5.4}$$

where k_0 is the number of lags used in defining matrix $\mathbf{M}$ in (4.3), and $K \geq 1$ is any fixed integer. (See also Fig.5.) Further simulation results, not reported explicitly, also conform with (5.4). This conjecture arises from the following observation: for $j > (k_0+1)r$, the j-th largest eigenvalue of $\widehat{\mathbf{M}}$ is predominately contributed by the term $\sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_{\varepsilon}(k) \widehat{\mathbf{\Sigma}}_{\varepsilon}(k)'$ which has a cluster of largest eigenvalues in the order of p^2/n^2, where $\widehat{\mathbf{\Sigma}}_{\varepsilon}(k)$ is the sample lag-k autocovariance matrix for $\boldsymbol{\varepsilon}_t$. See also Theorem 2(iii) in section 5.2.5 below.

(ii) Condition $h_n = p^{\delta} n^{-1/2} \to 0$ is very mild. It is always true if all the factors are strong (i.e. $\delta = 0$). In the case of weak factors with $\delta > 0$, it entails an upper bound order $p = o(n^{1/(2\delta)})$.

Table 1: Relative frequency estimates for $P(\widehat{r} = r)$ in the simulation with 200 replications

	n	50	100	200	400	800	1600	3200
$\delta = 0$	$p = 0.2n$	0.165	0.680	0.940	0.995	1	1	1
	$p = 0.5n$	0.410	0.800	0.980	1	1	1	1
	$p = 0.8n$	0.560	0.815	0.990	1	1	1	1
	$p = 1.2n$	0.590	0.820	0.990	1	1	1	1
$\delta = 0.5$	$p = 0.2n$	0.075	0.155	0.270	0.570	0.980	1	1
	$p = 0.5n$	0.090	0.285	0.285	0.820	0.960	1	1
	$p = 0.8n$	0.060	0.180	0.490	0.745	0.970	1	1
	$p = 1.2n$	0.090	0.180	0.310	0.760	0.915	1	1

For example, Corollary 1 implies that the estimator $\widehat{r}$ defined in (4.5) may work when $p = O(n)$ and the factors are not too weak in the sense that $\delta < 1/2$.

(iii) The errors in estimating eigenvalues are in the order of $p^{2-\delta}n^{-1/2}$ or p^2n^{-1}, and both do not necessarily converge to 0. However since

$$\frac{\widehat{\lambda}_j}{|\widehat{\lambda}_i - \lambda_i|} = O_P(p^\delta n^{-1/2}) = O_P(h_n) = o_P(1), \quad \text{for any } 1 \le i \le r \text{ and } r < j \le p,$$

the estimation errors for the zero-eigenvalues is asymptotically of an order of magnitude smaller than those for the non-zero eigenvalues.

5.2.4 Simulation

To illustrate the above asymptotic properties, we report some simulation results. We set in model (3.1) $r = 3$, $n = 50, 100, 200, 400, 800, 1600$ and 3200, and $p = 0.2n$, $0.5n$, $0.8n$ and $1.2n$. All the $p \times r$ elements of $\mathbf{A}$ are generated independently from the uniform distribution on the interval $[-1, 1]$ first, and we then divide each of them by $p^{\delta/2}$ to make all 3 factors of the strength δ; see (5.2). We generate factor $\mathbf{x}_t$ from a 3×1 vector-AR(1) process with independent $N(0, 1)$ innovations and the diagonal autoregressive coefficient matrix with 0.6, -0.5 and 0.3 as the main diagonal elements. We let ε_t in (3.1) consist of independent $N(0, 1)$ components and they are also independent across t. We set $k_0 = 1$ in (4.3) and (4.4). For each setting, we replicate the simulation 200 times.

Table 1 reports the relative frequency estimates for the probability $P(\widehat{r} = r) = P(\widehat{r} = 3)$ with $\delta = 0$ and 0.5. When the factors are strong (i.e. $\delta = 0$), the estimation for r illustrates the 'blessing of dimensionality' at its clearest. For example, when the sample size $n = 100$, the relative frequencies for $\widehat{r} = r$ are, respectively, 0.68, 0.8, 0.815 and 0.82 for $p = 20, 50$, 80 and 120. The improvement is due to the increased information on r from the added components of $\mathbf{y}_t$ when p increases. When $\delta = 0.5$, the columns of $\mathbf{A}$ are p-vectors with the norm $p^{0.25}$ (see (5.2)). Hence we may think that many elements of $\mathbf{A}$ are now effectively 0. The increase of the information on the factors is coupled with the increase of 'noise' when p increases. Indeed, Table 1 shows that

when factors are weak as $\delta = 0.5$, the estimation for r is not necessarily improved as p increases.

Our simulation results, not reported here to save the space, also show that the estimated eigenvalues are not consistent when n and p increase together, although the estimations errors for the zero-eigenvalues are much smaller than those for the non-zero eigenvalues. Those results conform with Theorem 1. See also Remark 2(iii).

We also experiment with a setting with two strong factors (with $\delta = 0$) and one weak factor (with $\delta = 0.5$). Then the ratio-based estimator $\widehat{r}$ tends to take values 2, picking up the two strong factors only. However Fig.6 indicates that the information on the third weak factor is not lost. In fact, $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ tends to take the second smallest value at $i = 3$. In this case a two-step estimation procedure should be employed in order to identify the number of factors correctly; see section 6 below.

5.2.5 Improved rates for the estimated eigenvalues

The rates in Proposition 3 and Theorem 1 can be further improved with the use of the random matrix theory, if we are prepared to entertain some additional conditions on $\boldsymbol{\varepsilon}_t$ in model (3.1). Such an improvement is relevant as the condition that $h_n = p^{\delta} n^{-1/2} \to 0$, required in Theorem 1, is sometimes unnecessary. For example in Table 1, the ratio-based estimator $\widehat{r}$ works perfectly well when $\delta = 0.5$ and n is sufficiently large (e.g. $n \geq 800$), even though $h_n = (p/n)^{1/2} \not\to 0$.

Now we introduce some additional conditions on $\boldsymbol{\varepsilon}_t$.

C7. Let ε_{jt} denote the j-th component of $\boldsymbol{\varepsilon}_t$. Then ε_{jt} are independent for different t and j, and have mean 0 and common variance $\sigma^2 < \infty$.

C8. The distribution of each ε_{jt} is symmetric. Furthermore $E(\varepsilon_{jt}^{2k+1}) = 0$, and $E(\varepsilon_{jt}^{2k}) \leq (\tau k)^k$ for all $1 \leq j \leq p$ and $t, k \geq 1$, where $\tau > 0$ is a constant independent of j, t, k.

The moment condition $E(\varepsilon_{jt}^{2k}) \leq (\tau k)^k$ in C8 implies that ε_{jt} are sub-Gaussian. When all components of $\{\boldsymbol{\varepsilon}_t\}$ are independent $N(0, \sigma^2)$, C7 and C8 holds. See also conditions (i') - (iv') of Péché (2009).

Theorem 2 *Let conditions C1-C8 hold, $\ell_n \equiv p^{\delta/2} n^{-1/2} \to 0$ and $n = O(p)$. Then as $n \to \infty$, it holds that*

$$\|\widehat{\mathbf{A}} - \mathbf{A}\| = O_P(\ell_n) \xrightarrow{P} 0. \tag{5.5}$$

Furthermore, the following assertions also hold.

(i) $|\widehat{\lambda}_j - \lambda_j| = O_P(p^{2-3\delta/2} n^{-1/2}) = O_P(p^{2-2\delta} \ell_n)$ *for* $j = 1, \cdots, r$,

(ii) $\widehat{\lambda}_j = O_P(p^{2-\delta} n^{-1}) = O_P(p^{2-2\delta} \ell_n^2)$ *for* $j = r+1, \cdots, (k_0+1)r$,

(iii) $\widehat{\lambda}_j = O_P(p^2 n^{-2}) = O_P(p^{2-2\delta} \ell_n^4)$ *for* $j = (k_0+1)r + 1, \cdots, p$.

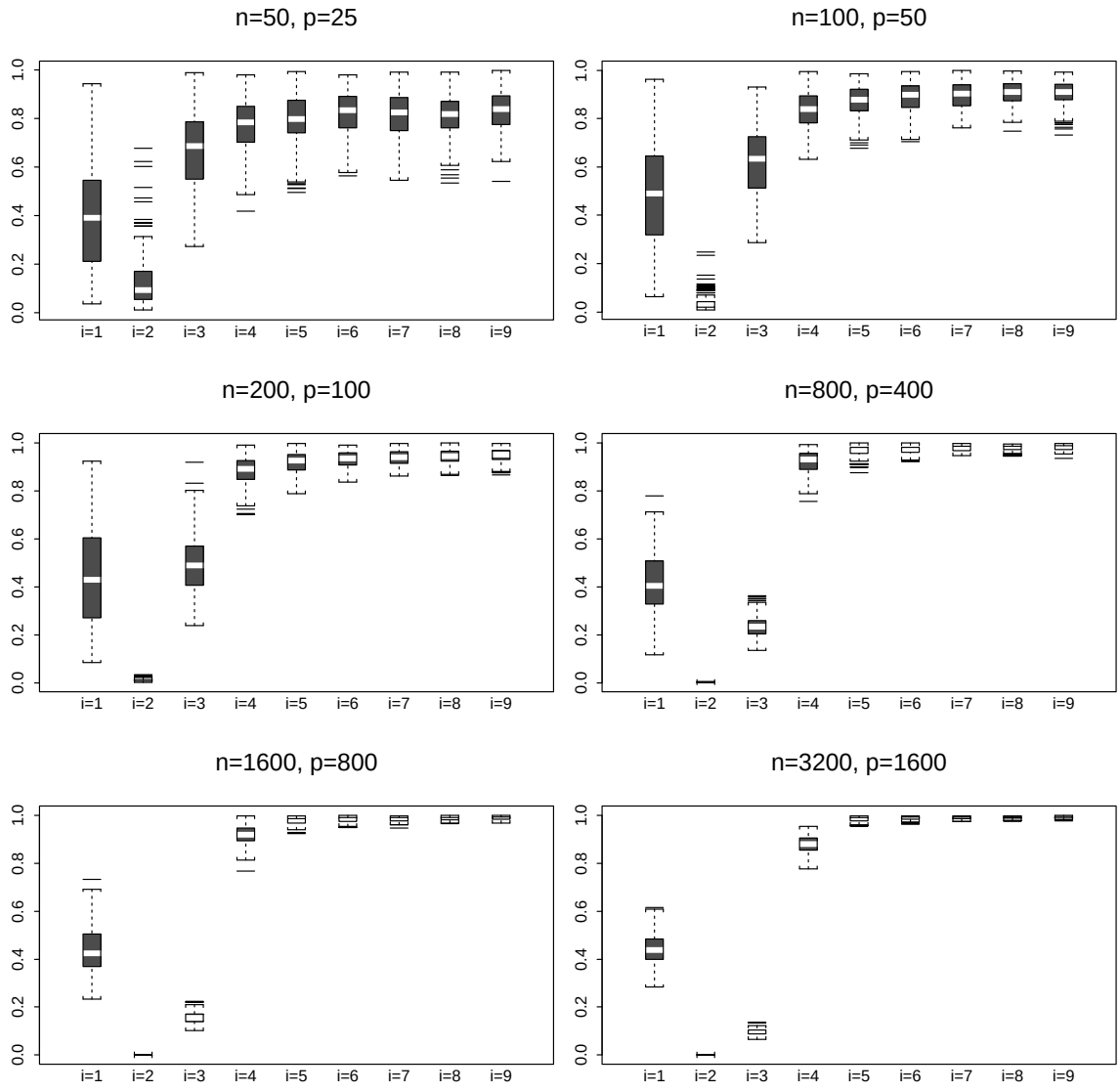

Figure 6: *Boxplots for the ratios* $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ *with two strong factors* ($\delta = 0$) *and one weak factor* ($\delta = 0.5$), *and* $r = 3$, $p = n/2$.

The proof of Theorem 2 is given in the Appendix. The corollary below follows immediately from the above theorem and the fact that $\lambda_j \asymp p^{2-2\delta}$ for $j = 1, \cdots, r$ (see condition C5 and Remark 1(i)).

Corollary 2 *Under the condition of Theorem 2, it holds that*

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \asymp 1 \;\; j = 1, \cdots, r-1, \quad \textit{and} \;\; \widehat{\lambda}_{r+1}/\widehat{\lambda}_r = O_P(p^{\delta} n^{-1}).$$

Remark 3. (i) By comparing with Proposition 3, the convergence rate for $\widehat{\mathbf{A}}$ in (5.5) is faster by a factor of $\ell_n/h_n = p^{-\delta/2}$.

(ii) The rates for the errors in estimating λ_j in Theorem 2 are improved over those in Theorem 1. More precisely the improvement for non-zero λ_j is by a factor $p^{-\delta/2}$, and for zero-

eigenvalues is by a factor $p^{-\delta}$ at least. However, those estimation errors themselves may still diverge, as illustrated in the simulation in section 5.2.4.

(iii) Theorem 2(iii) is an interesting consequence of the random matrix theory. The key message here is as follows: For eigenvalues corresponding purely to the matrix $\sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_\varepsilon(k)\widehat{\mathbf{\Sigma}}_\varepsilon(k)'$, which is a part of $\widehat{\mathbf{M}}$ in (4.4), where $\widehat{\mathbf{\Sigma}}_\varepsilon(k)$ is the sample lag-k autocovariance matrix for $\{\boldsymbol{\varepsilon}_t\}$, their magnitudes adjusted for $p^{2-2\delta}$ converge at a super-fast rate. In particular, when all the factors are strong (i.e. $\delta = 0$), the convergence rate is n^{-2}. Such a super convergence rate never occurs when p is fixed.

(iv) Condition $\ell_n \to 0$ is weaker than condition $h_n \to 0$. For example when $p \asymp n$, this condition is implied by the condition $\delta \in [0, 1)$.

6 Two-step estimation

In this section, we outline a two-step estimation procedure. We will show that it is superior than the one-step procedure presented in section 4.2 for the determination of the number of factors in the presence of the factors with different degrees of strength. (See Example 2 in section 7 and also Fig. 6.) Peña and Poncela (2006) described a similar procedure to improve the estimation for factor loading matrices in the presence of small eigenvalues, although they gave no theoretical underpinning on why and when such a procedure is advantageous.

Consider model (3.1) with r_1 strong factors with strength $\delta_1 = 0$ and r_2 weak factors with strength $\delta_2 > 0$, where $r_1 + r_2 = r$. Now (3.1) may be written as

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t = \mathbf{A}_1\mathbf{x}_{1t} + \mathbf{A}_2\mathbf{x}_{2t} + \boldsymbol{\varepsilon}_t, \tag{6.1}$$

where $\mathbf{x}_t = (\mathbf{x}_{1t}' \ \mathbf{x}_{2t}')'$, $\mathbf{A} = (\mathbf{A}_1 \ \mathbf{A}_2)$ with $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$, $\mathbf{x}_{1t}$ consists of r_1 strong factors, and $\mathbf{x}_{2t}$ consists of r_2 weak factors.

To present the two-step estimation procedure clearly, let us assume that we know r_1 and r_2 first. Using the method in section 4.2, we first obtain the estimator $\widehat{\mathbf{A}} \equiv (\widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2)$ for the factor loading matrix $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$, where the columns of $\widehat{\mathbf{A}}_1$ are the r_1 orthonormal eigenvectors of $\widehat{\mathbf{M}}$ corresponding to its r_1 largest eigenvalues. In practice we may identify r_1 using, for example, the ratio-based estimation method (4.5); see Fig. 6. We adopt the second-step estimation as follows. Let

$$\mathbf{y}_t^* = \mathbf{y}_t - \widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1'\mathbf{y}_t \tag{6.2}$$

for all t. We perform the same estimation for data $\{\mathbf{y}_t^*\}$ now, and obtain the $p \times r_2$ estimated factor loading matrix $\widetilde{\mathbf{A}}_2$ for the r_2 weak factors. Combining the two estimators together, we obtain the final estimator for $\mathbf{A}$ as

$$\widetilde{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \ \widetilde{\mathbf{A}}_2). \tag{6.3}$$

Theorem 3 below presents the convergence rates for both the one-step estimator $\widehat{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2)$ and the two-step estimator $\widetilde{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \ \widetilde{\mathbf{A}}_2)$. It shows that $\widetilde{\mathbf{A}}$ converges to $\mathbf{A}$ at a faster rate than

$\widehat{\mathbf{A}}$. The results are established with known r_1 and r_2. In practice we estimate r_1 and r_2 using the ratio based estimators. See also Corollary 3 below. We introduce some regularity conditions first. Let $\mathbf{\Sigma}_{12}(k) = \mathrm{Cov}(\mathbf{x}_{1,t+k}, \mathbf{x}_{2t})$, $\mathbf{\Sigma}_{21}(k) = \mathrm{Cov}(\mathbf{x}_{2,t+k}, \mathbf{x}_{1t})$, $\mathbf{\Sigma}_i(k) = \mathrm{Cov}(\mathbf{x}_{i,t+k}, \mathbf{x}_{it})$ and $\mathbf{\Sigma}_{i\epsilon}(k) = \mathrm{Cov}(\mathbf{x}_{i,t+k}, \boldsymbol{\varepsilon}_t)$ for $i = 1, 2$.

C5'. For $i = 1, 2$, $1 \le k \le k_0$,

$$\|\mathbf{\Sigma}_i(k)\| \asymp p^{1-\delta_i} \asymp \|\mathbf{\Sigma}_i(k)\|_{\min}, \quad \|\mathbf{\Sigma}_{21}(k)\| \asymp \|\mathbf{\Sigma}_{21}(k)\|_{\min}, \quad \|\mathbf{\Sigma}_{12}(k)\| = O(p^{1-\delta_2/2}).$$

C6'. $\mathrm{Cov}(\mathbf{x}_t, \varepsilon_s) = 0$ for any t, s.

The condition on $\mathbf{\Sigma}_i(k)$ in C5' is an analogue to condition C5. See Remark 1(i) in section 5.2.2 for the background of those conditions. The order of $\|\mathbf{\Sigma}_{21}(k)\|_{\min}$ will be specified in the theorems below. The order of $\|\mathbf{\Sigma}_{12}(k)\|$ is not restrictive, since $p^{1-\delta_2/2}$ is the largest possible order when $\delta_1 = 0$. See also the discussion in Remark 1(ii). Condition C6' replaces condition C6. Here we impose a strong condition $\mathbf{\Sigma}_{i\epsilon}(k) = 0$ to highlight the benefits of the two-step estimation procedure. See Remark 4(ii) below. Put

$$\mathbf{W}_i = (\mathbf{\Sigma}_i(1), \cdots, \mathbf{\Sigma}_i(k_0)), \qquad \mathbf{W}_{21} = (\mathbf{\Sigma}_{21}(1), \cdots, \mathbf{\Sigma}_{21}(k_0)).$$

Theorem 3 *Let conditions C1-C4, C5', C6', C7 and C8 hold. Let $n = O(p)$ and $\kappa_n \equiv p^{\delta_2/2} n^{-1/2} \to 0$, as $n \to \infty$. Then it holds that*

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\| = O_P(n^{-1/2}), \quad \|\widetilde{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(\kappa_n) = \|\widetilde{\mathbf{A}} - \mathbf{A}\|.$$

Furthermore,

$$\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(\nu_n) = \|\widehat{\mathbf{A}} - \mathbf{A}\|$$

if, in addition, $\nu_n \to 0$, where

$$\textit{where } \nu_n = \begin{cases} p^{\delta_2}\kappa_n, & \textit{if } \|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2}); \\ p^{(2c-1)\delta_2}\kappa_n, & \textit{if } \|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2} \textit{ for } 1/2 \le c < 1, \textit{ and} \\ & \|\mathbf{W}_1\mathbf{W}_{21}'\| \le q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\| \textit{ for } 0 \le q < 1. \end{cases}$$

Note that $\kappa_n/\nu_n \to 0$. Theorem 3 indicates that between $\mathbf{A}_1$ and $\mathbf{A}_2$, the latter is more difficult to estimate, and the convergence rate of an estimator for $\mathbf{A}$ is determined by such a rate for $\mathbf{A}_2$. This is intuitively understandable as the coefficient vectors for weak factors effectively contain many zero-components; see (3.4) and (5.2). Therefore a non-trivial proportion of the components of $\mathbf{y}_t$ may contain little information on each weak factor. When $\|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}$, $\|\mathbf{W}_2\|$ is dominated by $\|\mathbf{W}_{21}\|_{\min}$. The condition $\|\mathbf{W}_1\mathbf{W}_{21}'\| \le q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\|$ for $0 \le q < 1$ is imposed to control the behaviour of the (r_1+1)-th to the r-th largest eigenvalues of $\mathbf{M}$ under this situation. If this is not valid, those eigenvalues can become very small and give a bad estimator for $\mathbf{A}_2$, and thus $\mathbf{A}$. Under this condition, the structure of the autocovariance for the strong

factors, and the structure of the cross-autocovariance between the strong and weak factors, are not similar.

Recall that λ_j and $\widehat{\lambda}_j$ are the j-th largest eigenvalue of, respectively, $\mathbf{M}$ defined in (4.3) and $\widehat{\mathbf{M}}$ defined in (4.4). We define matrices $\mathbf{M}^*$ and $\widehat{\mathbf{M}}^*$ in the same manners as $\mathbf{M}$ and $\widehat{\mathbf{M}}$ but with $\{\mathbf{y}_t\}$ replaced by $\{\mathbf{y}_t^*\}$ (see (6.2)), and denote by λ_j^* and $\widehat{\lambda}_j^*$ the j-th largest eigenvalue of, respectively, $\mathbf{M}^*$ and $\widehat{\mathbf{M}}^*$. To gain the appreciation on how the ratio-based estimator (4.5) works in the presence of factors with different degrees of strength, we present some asymptotic properties for the estimated eigenvalues of $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}^*$ first.

Theorem 4 *Under the same conditions and notations of Theorem 3, the following assertions hold.*

(i) For $j = 1, \cdots, r_1$, $|\widehat{\lambda}_j - \lambda_j| = O_P(p^2 n^{-1/2})$ *and* $\lambda_j \asymp p^2$.

(ii) For $j = r_1 + 1, \cdots, r$, $|\widehat{\lambda}_j - \lambda_j| = O_P(p^2(n^{-1/2} + \nu_n^2))$ *and*

$$\lambda_j \asymp \begin{cases} p^{2-2\delta_2}, & \text{if } \|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2}); \\ p^{2-2c\delta_2}, & \text{if } \|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2} \text{ for } 1/2 \le c < 1, \text{ and} \\ & \|\mathbf{W}_1 \mathbf{W}_{21}'\| \le q \|\mathbf{W}_1\|_{\min} \|\mathbf{W}_{21}\| \text{ for } 0 \le q < 1, \end{cases}$$

provided $\nu_n \to 0$.

(iii) For $j = r + 1, \cdots, p$, $\widehat{\lambda}_j = O_P(p^2 \nu_n^2)$, *provided* $\nu_n \to 0$.

(iv) For $j = 1, \cdots, r_2$, $|\widehat{\lambda}_j^* - \lambda_j^*| = O_P(p^{2-2\delta_2}\kappa_n)$, $\lambda_j^* \asymp p^{2-2\delta_2}$.

(v) For $j = r_2 + 1, \cdots, p$, $\widehat{\lambda}_j^* = O_P(p^{2-2\delta_2}\kappa_n^2)$.

(vi) For $j = (k_0 + 1)r + 1, \cdots, p$, $\widehat{\lambda}_j$, $\widehat{\lambda}_j^* = O_P(p^2 n^{-2}) = O_P(p^{2-2\delta_2}\kappa_n^4)$.

Corollary 3 follows from Theorem 4. It presents the asymptotic properties for the ratios of the estimated eigenvalues.

Corollary 3 *Let all the conditions in Theorem 4 hold. Then the following assertions hold.*

(i) $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i \asymp 1$ *for* $1 \le i < r_1$ *and* $r_1 < i < r$, *and* $\widehat{\lambda}_{j+1}^*/\widehat{\lambda}_j^* \asymp 1$ *for* $1 \le 1 < r_2$.

(ii) $\widehat{\lambda}_{r+1}/\widehat{\lambda}_r \xrightarrow{P} 0$. *Furthermore* $\widehat{\lambda}_{r_1+1}/\widehat{\lambda}_{r_1} = o_P(\widehat{\lambda}_{r+1}/\widehat{\lambda}_r) \xrightarrow{P} 0$.

(ii) $\widehat{\lambda}_{r_2+1}^*/\widehat{\lambda}_{r_2}^* = o_P(\widehat{\lambda}_{r+1}/\widehat{\lambda}_r) \xrightarrow{P} 0$.

Remark 4. (i) Corollary 3 implies that the one-step estimation is likely to lead to $\widehat{r} = r_1$; only picking up the r_1 strong factors, while the two-step estimation will be able to identify the additional r_2 factors. Unfortunately we are unable to establish the asymptotic properties for $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ for $i > r$, and $\widehat{\lambda}_{j+1}^*/\widehat{\lambda}_j^*$ for $j > r_2$, though we believe that conjectures similar to (5.4) continue to hold.

(ii) When $\delta_1 > 0$ and/or the cross-autocovariances between different factors and the noise are stronger, the similar and more complex results can be established via more involved algebra in the proofs.

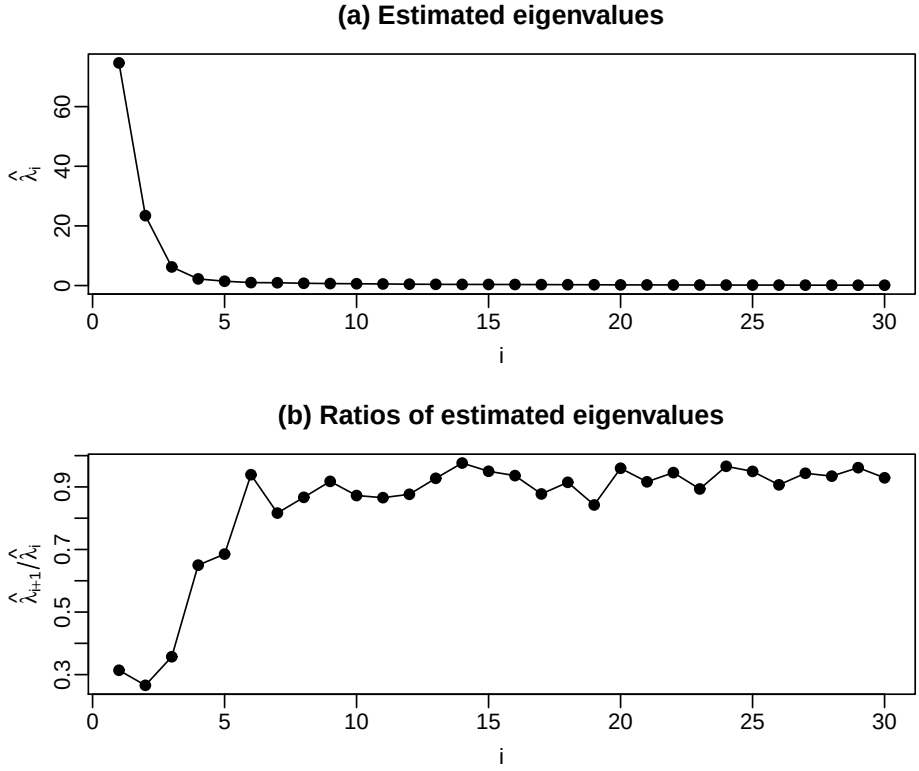

Figure 7: Plots of the eigenvalues and the ratios of eigenvalues of $\widehat{\mathbf{M}}$ for Example 1.

7 Real data examples

Example 1. First we analyze the daily returns of 123 stocks in the period 2 January 2002 — 11 July 2008. Those stocks were selected among those included in the S&P500 and were traded everyday during the period. The returns were calculated in percentages based on the daily close prices. We have in total $n = 1642$ observations with $p = 123$. We apply the eigenanalysis to the matrix $\widehat{\mathbf{M}}$ defined in (4.4) with $k_0 = 5$. The obtained eigenvalues (in descending order) and their ratios are plotted in Fig.7. It is clear that the ratio-based estimator (4.5) leads to $\widehat{r} = 2$, indicating 2 factors. Varying the value of k_0 between 1 and 100 in the definition of $\widehat{\mathbf{M}}$ leads to little change in the ratios $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$, and the estimate $\widehat{r} = 2$ remains unchanged. This is due to the fact that the return series exhibit little autocorrelation. An increase of the value k_0 effectively adds zero-matrices in $\widehat{\mathbf{M}}$; see (4.4). Fig.7(a) shows that $\widehat{\lambda}_i$ is close to 0 for all $i \geq 5$. Fig.7(b) indicates that the ratio $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ is close to 1 for all large i, which is in line with conjecture (5.4).

The first two panels of Fig.8 display the time series plots of the two component series of the estimated factors $\widehat{\mathbf{x}}_t$ defined as in (3.2). Their cross autocorrelations are presented in Fig.9. Although each of the two estimated factors shows little significant autocorrelation, there are some significant cross-correlations between the two series. Fig.10 presents the cross autocorrelations of the three residual series. Those three series are $\widehat{\boldsymbol{\gamma}}_j'\mathbf{y}_t$ for $j = 3, 4, 5$, where $\widehat{\boldsymbol{\gamma}}_j$ is the unit

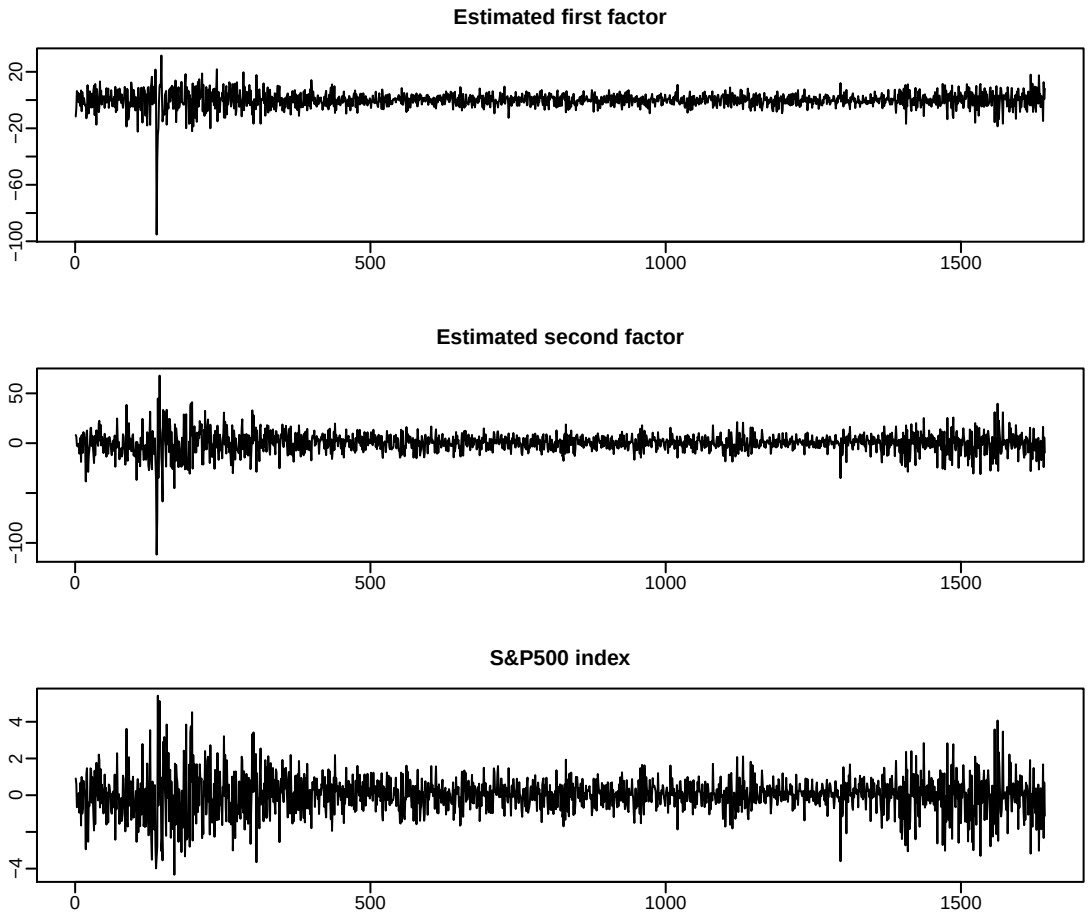

Figure 8: The time series plots of the two estimated factors and the return series of the S&P500 index in the same time period.

eigenvector of $\widehat{\mathbf{M}}$ corresponding to its j-th largest eigenvalue. If there were any serial correlations left in the data after extracting the two estimated factors, those correlations would most likely show up in the three selected residual series. Indeed there is little evidence for the existence of those correlations; see Fig.10.

Fig.7 suggests the existence of a third and weaker factor, though Fig.10 suggests otherwise. In fact $\widehat{\lambda}_3 = 6.231$ and $\widehat{\lambda}_4/\widehat{\lambda}_3 = 0.357$. Note that now $\widehat{\lambda}_j$ is not necessarily a consistent estimator for λ_j although $\widehat{\lambda}_{r+1}/\widehat{\lambda}_r \xrightarrow{P} 0$; see Theorem 1(ii) and Corollary 1. To investigate this further, we apply the two-step estimation procedure presented in section 6. By subtracting the two estimated factors from the above, we obtain the new data $\mathbf{y}_t^*$ (see (6.3)). We then calculate the eigenvalues and their ratios of the matrix $\widehat{\mathbf{M}}^*$. The minimum value of the ratios is $\widehat{\lambda}_2^*/\widehat{\lambda}_1^* = 0.667$, which is closely followed by $\widehat{\lambda}_3^*/\widehat{\lambda}_2^* = 0.679$ and $\widehat{\lambda}_4^*/\widehat{\lambda}_3^* = 0.744$. There is no evidence to suggest that $\widehat{\lambda}_2^*/\widehat{\lambda}_1^* \to 0$; see Corollary 3. This reinforces our choice $\widehat{r} = 2$.

With p as large as 123, it is difficult to gain insightful interpretation on the estimated factors

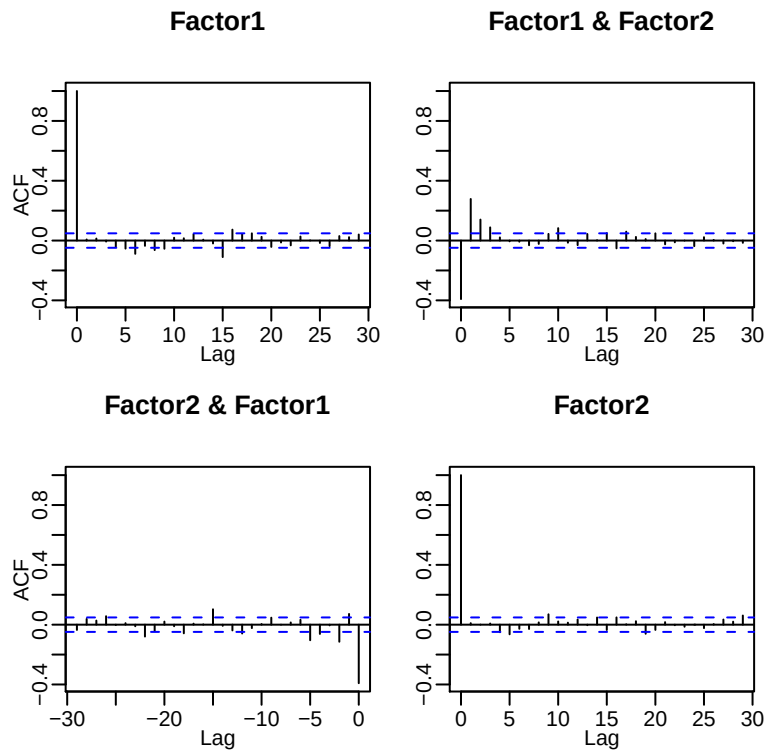

Figure 9: The cross autocorrelations of the two estimated factors for Example 1.

by looking through the coefficients in $\widehat{\mathbf{A}}$ (see (3.2)). To link our fitted factor model with some classical asset pricing theory in finance, we wonder if the market index (i.e. the S&P500 index) is a factor in our fitted model, or more precisely, if it can be written as a linear combination of the two estimated factors. When this is true, $\mathbf{Pu} = 0$, where $\mathbf{u}$ is the 1642×1 vector consisting of the returns of the S&P500 index over the same time period, and $\mathbf{P}$ denotes the projection matrix onto the orthogonal compliment of the linear space spanned by the two component series $\widehat{\mathbf{x}}_t$, which is a 1640-dimensional subspace in R^{1642}. This S&P500 return series is plotted together with the two component series $\widehat{\mathbf{x}}_t$ in Fig.8. It turns out that $||\mathbf{Pu}||^2$ is not exactly 0 but $||\mathbf{Pu}||^2/||\mathbf{u}||^2 = 0.023$. i.e. the 97.7% of the S&P500 returns can be expressed as a linear combination of the two estimated factors. Thus our analysis suggests the following model for $\mathbf{y}_t$ — the daily returns of the 123 stocks:

$$\mathbf{y}_t = \mathbf{a}_1 u_t + \mathbf{a}_2 v_t + \boldsymbol{\varepsilon}_t,$$

where u_t denotes the return of the S&P500 on the day t, v_t is another factor, and $\boldsymbol{\varepsilon}_t$ is a 123×1 vector white noise process.

Example 2. We analyze a set of monthly average sea surface air pressure records (in Pascal) in January 1958 – December 2001 (i.e. 528 months in total) over a 10×44 grid in a range of 22.5° longitude and 110° latitude in the North Atlantic Ocean. Let $P_t(u, v)$ denote the air pressure in the t-th month at the location (u, v), where $u = 1, \cdots, 10,\ v = 1, \cdots, 44$ and $t = 1, \cdots, 528$.

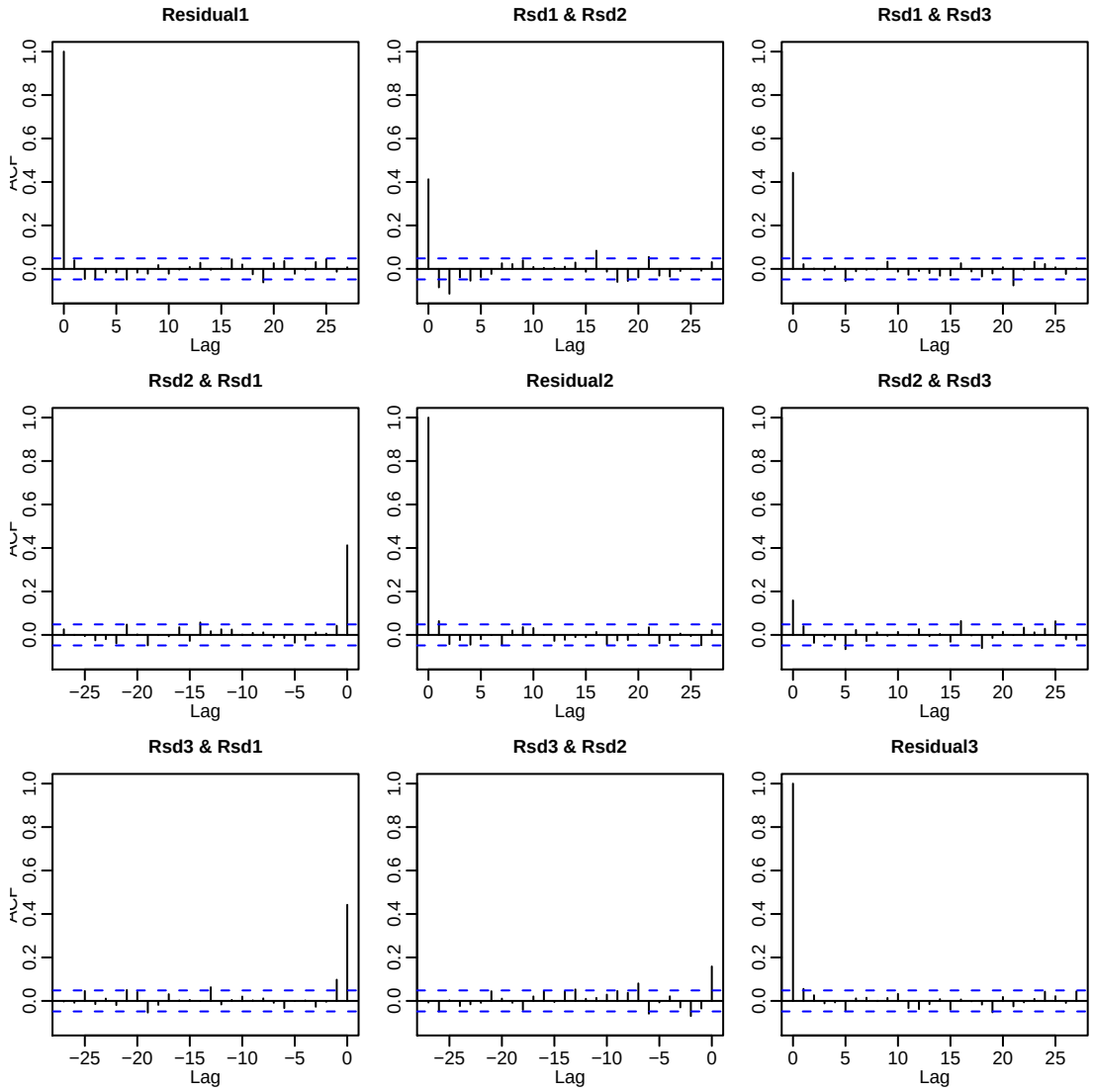

Figure 10: The cross autocorrelations of three residual series for Example 1.

Fig.11 displays the time series plots of those data at three different locations. We first subtract each data point by the monthly mean over the 44 years at its location: $\frac{1}{44}\sum_{i=1}^{44} P_{12(i-1)+j}(u,v)$, where $j = 1, \cdots, 12$, representing the 12 different months over a year. We then line up the new data over $10 \times 44 = 440$ grid points as a vector $\mathbf{y}_t$, so that $\mathbf{y}_t$ is a p-variate time series with $p = 440$. We have $n = 528$ observations.

To fit the factor model (3.1) to $\mathbf{y}_t$, we calculate the eigenvalues and the eigenvectors of the matrix $\widehat{\mathbf{M}}$ defined in (4.4) with $k_0 = 5$. Let $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \cdots$ denote the eigenvalues of $\widehat{\mathbf{M}}$. The ratios $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ are plotted against i in the top panel of Fig.12 which indicates the ratio-based estimate for the number of factor is $\widehat{r} = 1$; see (4.5). However the second smallest ratio is $\widehat{\lambda}_4/\widehat{\lambda}_3$, This suggests that there may exist two weaker factors in addition; see Corollary 3(ii) and also

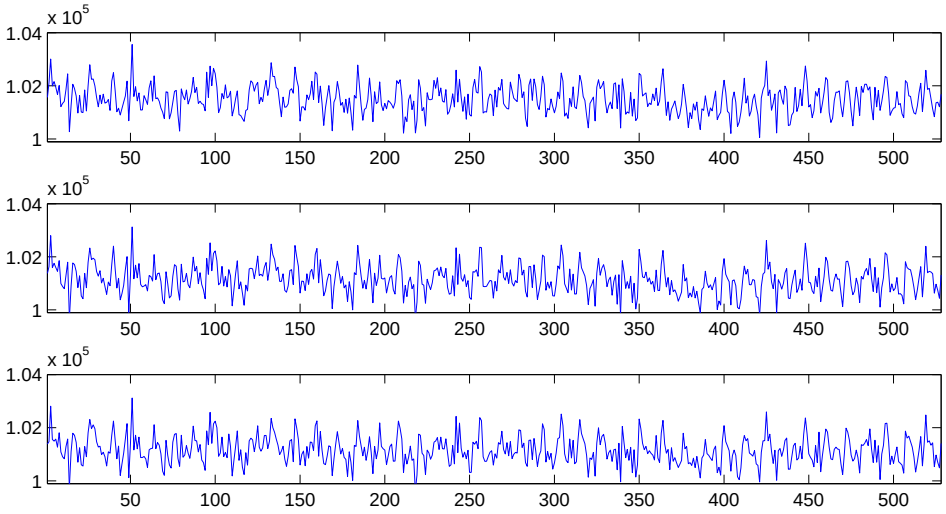

Figure 11: *Time series plot the sea surface air pressure data at the three locations (from top to bottom):* $(u, v) = (1, 1),\ (3, 4)$ *and* $(5, 6)$.

Fig.6. We adopt the two-step estimation procedure presented in section 6 to identify the factors of different strength. By removing the factor corresponding to the largest eigenvalue of $\widehat{\mathbf{M}}$, the resulting 'residuals' are denoted as $\mathbf{y}_t^*$; see (6.2). Now we repeat the factor modelling for data $\mathbf{y}_t^*$, and plot the ratios of eigenvalues of matrix $\widehat{\mathbf{M}}^*$ in the second panel of Fig.12. It shows clearly the minimum value at 2, indicating further two (weaker) factors. Combining the above two steps together, we set $\widehat{r} = 3$ in the fitted model.

We repeated the above calculation with $k_0 = 1$ in (4.4). We still find three factors with the two-step procedure, and the estimated factors series are very similar to the case when $k_0 = 5$. This is consistent with the simulation results in Lam, Yao and Bathia (2010), where they showed empirically that the estimated factor models is not sensitive to the choice of k_0.

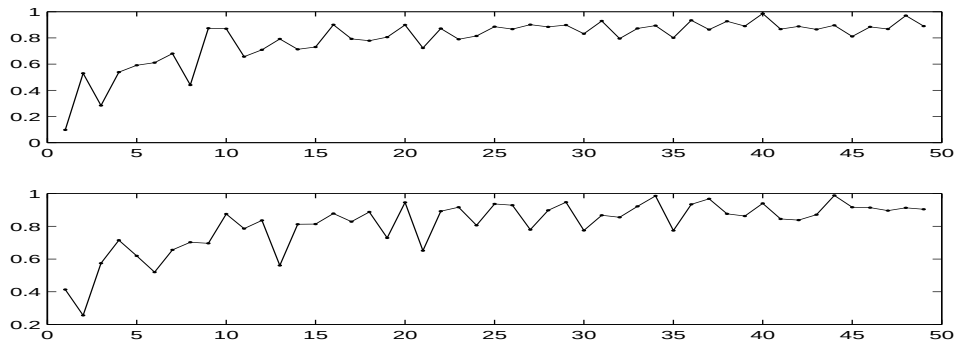

Figure 12: *Plots of* $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ *— the ratio of the eigenvalues of* $\widehat{\mathbf{M}}$ *(the top panel) and* $\widehat{\mathbf{M}}^*$ *(the bottom panel), against* i*, for Example 2.*

We present the time series plots for the three estimated factors $\widehat{\mathbf{x}}_t = \widetilde{\mathbf{A}}'\mathbf{y}_t$ in Fig.13, where $\widetilde{\mathbf{A}}$ is a 440×3 matrix with the first column being the unit eigenvector of $\widehat{\mathbf{M}}$ corresponding to its largest eigenvalue, and the other two columns being the orthonormal eigenvectors of $\widehat{\mathbf{M}}^*$ corresponding to its two largest eigenvalues; see (6.3) and also (3.2). They collectively account for 85.3% of the total variation in $\mathbf{y}_t$ which has 440 component series. In fact each of the three factors accounts for, respectively, 57.5%, 18.2% and 9.7% of the total variation of $\mathbf{y}_t$. Fig.14 depicts the the factor

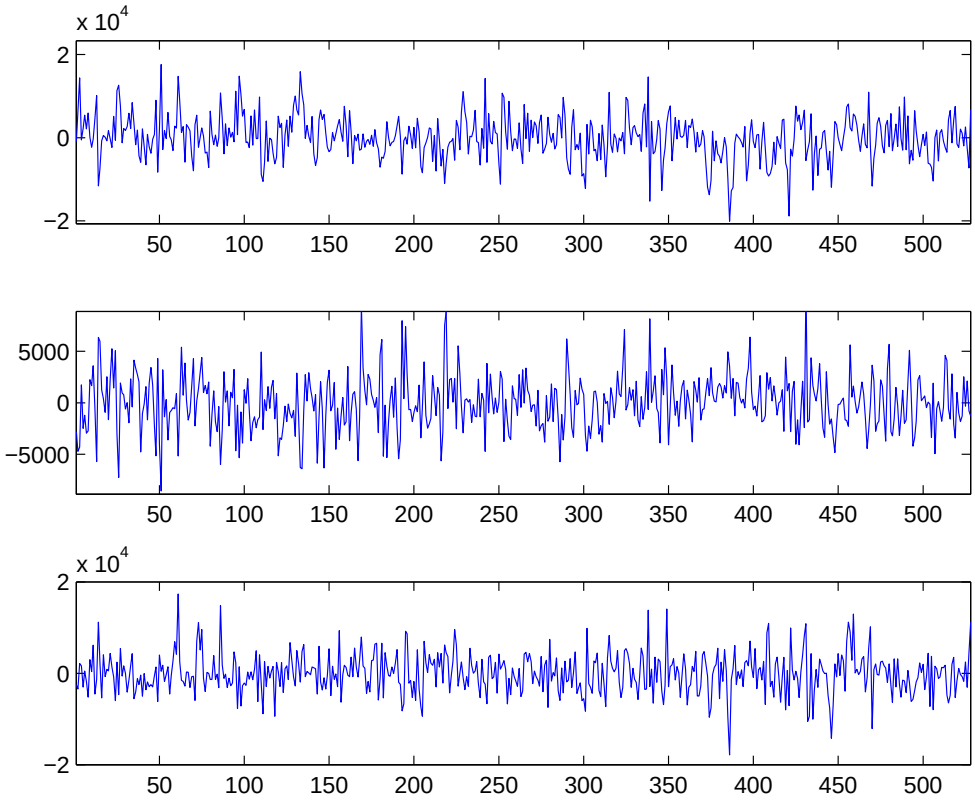

Figure 13: *Time series plot of the three estimated factors for Example 2.*

loading surfaces of the three factors. Some interesting regional patterns are observed from those plots. For example, the first factor is the main driving force for the dynamics in the north and especially the northeast. The second factor influences the dynamics in the east and the west in the opposite directions, and has little impact in the narrow void between them. The third factor impacts mainly the dynamics of the southeast region. We also notice that none of those factors can be seen as idiosyncratic components as each of them affects quite a large number of locations.

Fig.15 presents the sample cross-correlations for the three estimated factors. It shows significant, though small, auto-correlations or cross-correlations at some non-zero lags. Fig.16 is the sample cross-correlations for three residuals series selected from three locations for which one is far apart from the other two spatially, showing little autocorrelations at non-zero lags. This

indicates that our approach is capable to identify the factors based on serial correlations.

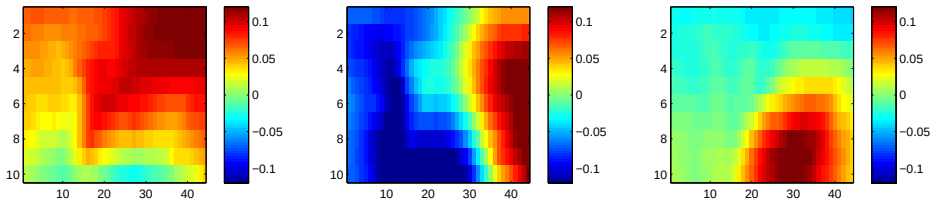

Figure 14: *Factor loading surface of the 1st, 2nd and 3rd factors (from left to right) for Example 2.*

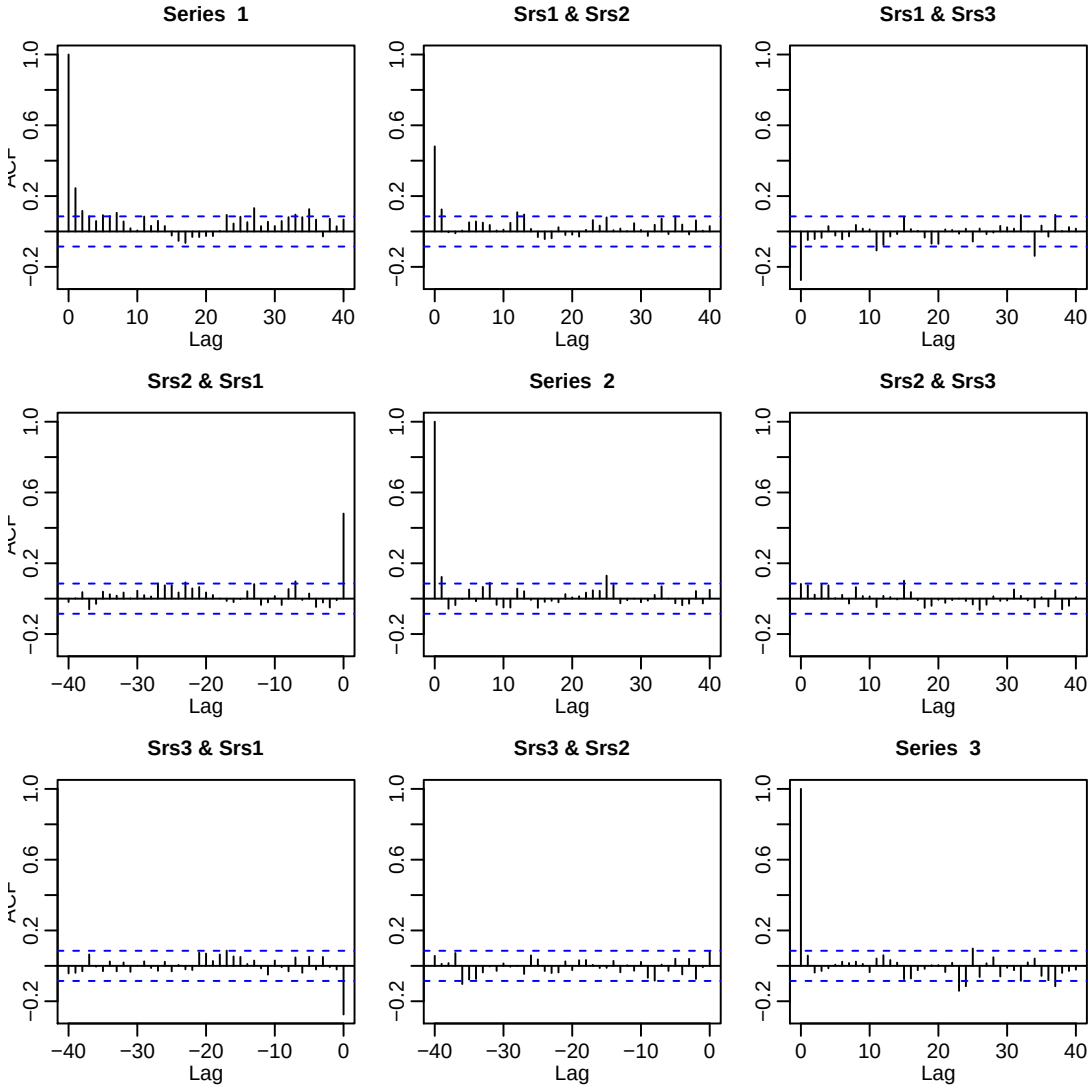

Figure 15: *Example 2: Sample cross-correlation functions for the three estimated factors.*

Finally we note that the BIC method of Bai and Ng (2002) yields the estimate $\widehat{r} = n = 528$ for this particular date set. We suspect that this may be due to the fact that Bai and Ng (2002)

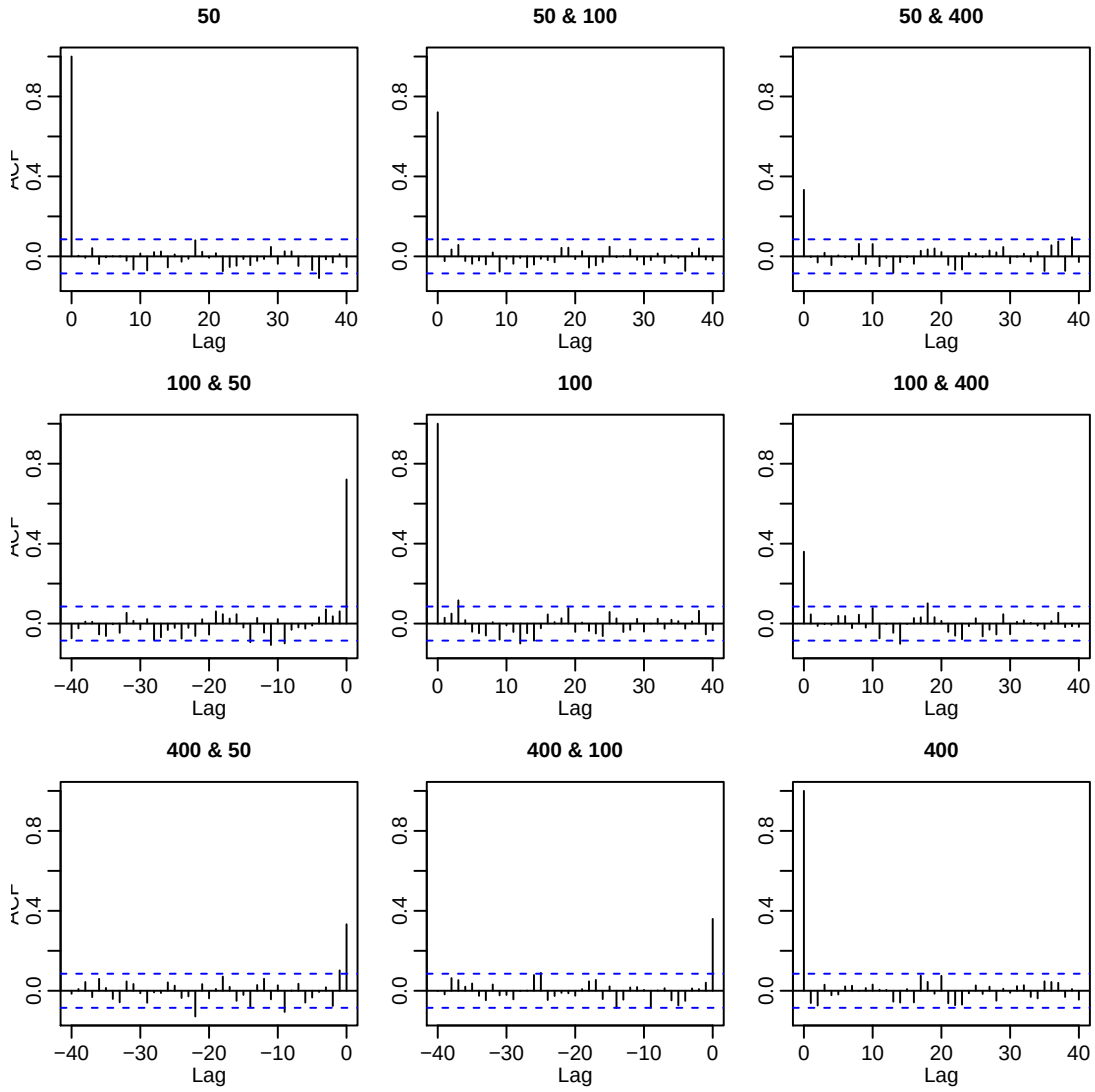

Figure 16: *Example 2: Sample cross-correlation functions for 3 residual series. 50 represents grid position (10, 5), 100 for (10, 10) and 400 for (10, 40).*

requires all the eigenvalues of $\mathbf{\Sigma}_\varepsilon$ be uniformly bounded when $p \to \infty$. This may not be the case for this particular data set, as the nearby locations are strongly spatially correlated, which may lead to very large and also very small eigenvalues for $\mathbf{\Sigma}_\varepsilon$. Indeed for this data set, the three largest eigenvalues of $\widehat{\mathbf{\Sigma}}_\varepsilon$ are in the order of 10^6, and the three smallest eigenvalues are practically 0. Since the typical magnitude of $\widehat{\boldsymbol{\varepsilon}}_t$ is 10^2 from our analysis, we have done simulations (not shown here) showing that the typical largest eigenvalues for $\widehat{\mathbf{\Sigma}}_\varepsilon$, if $\{\boldsymbol{\varepsilon}_t\}$ is weakly correlated white noise, should be around 10^4 to 10^5, and the smallest around 10^2 to 10^3 when $p = 440$ and $n = 528$. Such a huge difference in the magnitude of the eigenvalues suggests strongly that the components of the white noise vector $\boldsymbol{\varepsilon}_t$ are strongly correlated. Our method does not require the uniform

boundedness of the eigenvalues of $\mathbf{\Sigma}_\varepsilon$.

8 Appendix

Proof of Theorem 1. We present some notational definitions first. We denote $\widehat{\lambda}_j, \widehat{\boldsymbol{\gamma}}_j$ the j-th largest eigenvalue of $\widehat{\mathbf{M}}$ and the corresponding orthonormal eigenvector, respectively. The corresponding population values are denoted by λ_j and $\mathbf{a}_j$ for the matrix $\mathbf{M}$. Hence $\widehat{\mathbf{A}} = (\widehat{\boldsymbol{\gamma}}_1, \cdots, \widehat{\boldsymbol{\gamma}}_r)$ and $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$. We also have

$$\lambda_j = \mathbf{a}_j' \mathbf{M} \mathbf{a}_j, \quad \widehat{\lambda}_j = \widehat{\boldsymbol{\gamma}}_j' \widehat{\mathbf{M}} \widehat{\boldsymbol{\gamma}}_j, \ j = 1, \cdots, p.$$

We show some intermediate results now. With condition (C3) and the fact that $\{\boldsymbol{\varepsilon}_t\}$ is white noise, we can easily see that (see also Lam, Yao and Bathia (2010)),

$$\widehat{\mathbf{\Sigma}}_x(k) - \mathbf{\Sigma}_x(k), \ \widehat{\mathbf{\Sigma}}_\epsilon(k) - \mathbf{\Sigma}_\epsilon(k), \ \widehat{\mathbf{\Sigma}}_{x\epsilon}(k) - \mathbf{\Sigma}_{x\epsilon}(k), \ \widehat{\mathbf{\Sigma}}_{\epsilon x}(k) = O_P(n^{-1/2}),$$

where $k = 0, 1, \cdots, k_0$. Then following the proof of Theorem 1 of Lam, Yao and Bathia (2010), we have the following for $k = 1, \cdots, k_0$:

$$\begin{aligned} \|\widehat{\mathbf{M}} - \mathbf{M}\| &= O_P(\|\mathbf{\Sigma}_y(k)\| \cdot \|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\|), \ \text{ where} \\ \|\mathbf{\Sigma}_y(k)\| &= O(p^{1-\delta}), \\ \|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\| &= O_P(p^{1-\delta} n^{-1/2} + p^{1-\delta/2} n^{-1/2} + \|\widehat{\mathbf{\Sigma}}_\epsilon(k)\|) \\ &= O_P(p^{1-\delta/2} n^{-1/2} + \|\widehat{\mathbf{\Sigma}}_\epsilon(k)\|). \end{aligned} \tag{8.1}$$

Now $\|\widehat{\mathbf{\Sigma}}_\epsilon(k)\| \le \|\widehat{\mathbf{\Sigma}}_\epsilon(k)\|_F = O_P(pn^{-1/2})$, where $\|\mathbf{M}\|_F = \text{trace}(\mathbf{M}\mathbf{M}')$ denotes the Frobenius norm of $\mathbf{M}$. Hence from (8.1),

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\| &= O_P(pn^{-1/2}), \ \text{ and} \\ \|\widehat{\mathbf{M}} - \mathbf{M}\| &= O_P(p^{1-\delta} \cdot pn^{-1/2}) = O_P(p^{2-\delta} n^{-1/2}). \end{aligned} \tag{8.2}$$

For the main proof, consider for $j = 1, \cdots, r$, the decomposition

$$\begin{aligned} \widehat{\lambda}_j - \lambda_j &= \widehat{\boldsymbol{\gamma}}_j' \widehat{\mathbf{M}} \widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j' \mathbf{M} \mathbf{a}_j \\ &= I_1 + I_2 + I_3 + I_4 + I_5, \ \text{ where} \\ I_1 &= (\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j)'(\widehat{\mathbf{M}} - \mathbf{M})\widehat{\boldsymbol{\gamma}}_j, \ I_2 = (\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j)'\mathbf{M}(\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j), \ I_3 = (\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j)'\mathbf{M}\mathbf{a}_j, \\ I_4 &= \mathbf{a}_j'(\widehat{\mathbf{M}} - \mathbf{M})\widehat{\boldsymbol{\gamma}}_j, \ I_5 = \mathbf{a}_j'\mathbf{M}(\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j). \end{aligned} \tag{8.3}$$

For $j = 1, \cdots, r$, since $\|\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j\| \le \|\widehat{\mathbf{A}} - \mathbf{A}\| = O_P(h_n)$ where $h_n = p^\delta n^{-1/2}$, and $\|\mathbf{M}\| \le \sum_{k=1}^{k_0} \|\mathbf{\Sigma}_y(k)\|^2 = O_P(p^{2-2\delta})$ by (8.1), together with (8.2) we have that

$$\|I_1\|, \ \|I_2\| = O_P(p^{2-2\delta} h_n^2), \quad \|I_3\|, \ \|I_4\|, \ \|I_5\| = O_P(p^{2-2\delta} h_n),$$

so that $|\widehat{\lambda}_j - \lambda_j| = O_P(p^{2-2\delta} h_n) = O_P(p^{2-\delta} n^{-1/2})$, which proves Theorem 1(i).

Now consider $j = r+1, \cdots, p$. Define

$$\widetilde{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_y(k)\mathbf{\Sigma}_y(k)', \quad \widehat{\mathbf{B}} = (\widehat{\boldsymbol{\gamma}}_{r+1}, \cdots, \widehat{\boldsymbol{\gamma}}_p), \quad \mathbf{B} = (\mathbf{a}_{r+1}, \cdots, \mathbf{a}_p).$$

Following the same proof of Theorem 1 of Lam, Yao and Bathia (2010), we can actually show that $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(h_n)$, so that $\|\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j\| \le \|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(h_n)$.

Noting $\lambda_j = 0$ for $j = r+1, \cdots, p$, consider the decomposition

$$\begin{aligned} &\widehat{\lambda}_j = \widehat{\boldsymbol{\gamma}}_j'\widehat{\mathbf{M}}\widehat{\boldsymbol{\gamma}}_j = K_1 + K_2 + K_3, \text{ where} \\ &K_1 = \widehat{\boldsymbol{\gamma}}_j'(\widehat{\mathbf{M}} - \widetilde{\mathbf{M}} - \widetilde{\mathbf{M}}' + \mathbf{M})\widehat{\boldsymbol{\gamma}}_j,\ K_2 = 2\widehat{\boldsymbol{\gamma}}_j'(\widetilde{\mathbf{M}} - \mathbf{M})(\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j), \\ &K_3 = (\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j)'\mathbf{M}(\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j). \end{aligned} \tag{8.4}$$

Using (8.2),

$$K_1 = \sum_{k=1}^{k_0} \|(\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k))'\widehat{\boldsymbol{\gamma}}_j\|^2 \le \sum_{k=1}^{k_0} \|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\|^2 = O_P(p^2 n^{-1}).$$

Similarly, using (8.1) and (8.2), and $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(h_n)$, we can show that

$$\begin{aligned} |K_2| &= O_P(\|\widetilde{\mathbf{M}} - \mathbf{M}\| \cdot \|\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j\|) = O_P(\|\widehat{\mathbf{M}} - \mathbf{M}\| \cdot \|\widehat{\mathbf{B}} - \mathbf{B}\|) = O_P(p^2 n^{-1}), \\ |K_3| &= O_P(\|\widehat{\mathbf{B}} - \mathbf{B}\|^2 \cdot \|\mathbf{M}\|) = O_P(p^{2-2\delta} h_n^2) = O_P(p^2 n^{-1}). \end{aligned}$$

Hence $\widehat{\lambda}_j = O_P(p^2 n^{-1})$, and the proof of the theorem completes. □

In the following, we use $\sigma_j(\mathbf{M})$ to denote the j-th largest singular value of a matrix $\mathbf{M}$, so that $\sigma_1(\mathbf{M}) = \|\mathbf{M}\|$. We use $\lambda_j(\mathbf{M})$ to denote the j-th largest eigenvalue of $\mathbf{M}$.

Proof of Theorem 2. The first part of the theorem is actually Theorem 2 of Lam, Yao and Bathia (2010). We prove the other parts of the theorem. Define $\mathbf{1}_k$ the column vector of k ones, and

$$\mathbf{E}_{r,s} = (\boldsymbol{\varepsilon}_r, \cdots, \boldsymbol{\varepsilon}_s) \text{ for } r \le s.$$

We now prove an important intermediate result. Since the asymptotic behaviour of the 3 sample means

$$\bar{\boldsymbol{\varepsilon}} = n^{-1}\mathbf{E}_{1,n}\mathbf{1}_n,\ (n-k)^{-1}\mathbf{E}_{1,n-k}\mathbf{1}_{n-k},\ (n-k)^{-1}\mathbf{E}_{k+1,n}\mathbf{1}_{n-k}$$

are exactly the same as k is finite and $\{\boldsymbol{\varepsilon}_t\}$ is stationary, in this proof we take the sample lag-k autocovariance matrix for $\{\boldsymbol{\varepsilon}_t\}$ to be

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_\epsilon(k) &= n^{-1}(\mathbf{E}_{k+1,n} - (n-k)^{-1}\mathbf{E}_{k+1,n}\mathbf{1}_{n-k}\mathbf{1}_{n-k}')(\mathbf{E}_{1,n-k} - (n-k^{-1}\mathbf{E}_{1,n-k}\mathbf{1}_{n-k}\mathbf{1}_{n-k}'))' \\ &= n^{-1}\mathbf{E}_{k+1,n}\mathbf{T}_{n-k}\mathbf{E}_{1,n-k}', \end{aligned}$$

where $\mathbf{T}_j = \mathbf{I}_j - j^{-1}\mathbf{1}_j\mathbf{1}_j'$. Then under conditions C7 and C8,

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_\epsilon(k)\| &\le \|n^{-1/2}\mathbf{E}_{k+1,n}\| \cdot \|\mathbf{T}_{n-k}\| \cdot \|n^{-1/2}\mathbf{E}_{1,n-k}\| \\ &= \lambda_1^{1/2}(n^{-1}\mathbf{E}_{k+1,n}'\mathbf{E}_{k+1,n}) \cdot \lambda_1^{1/2}(n^{-1}\mathbf{E}_{1,n-k}'\mathbf{E}_{1,n-k}) \\ &= O_P((1+(pn^{-1})^{1/2}) \cdot (1+(pn^{-1})^{1/2})) \\ &= O_P(pn^{-1}), \end{aligned} \tag{8.5}$$

where the second last line follows from Theorem 1.3 of Péché (2009) for the covariance matrices $n^{-1}\mathbf{E}_{k+1,n}'\mathbf{E}_{k+1,n}$ and $n^{-1}\mathbf{E}_{1,n-k}'\mathbf{E}_{1,n-k}$, and the last line follows from the assumption $n = O(p)$.

For the main proof of the theorem, note that (8.1) together with (8.5) implies

$$\|\widehat{\mathbf{M}} - \mathbf{M}\| = O_P(p^{1-\delta}(p^{1-\delta/2}n^{-1/2} + pn^{-1})) = O_P(p^{2-2\delta}(p^{\delta/2}n^{-1/2} + p^\delta n^{-1})) = O_P(p^{2-2\delta}\ell_n),$$

since $\ell_n = p^{\delta/2}n^{-1/2} = o(1)$. Note also that we have $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(\ell_n)$, similar to the proof of Theorem 1.

With these, for $j = 1, \cdots, r$, using decomposition (8.3), we have

$$|\widehat{\lambda}_j - \lambda_j| = O_P(\|\widehat{\mathbf{M}} - \mathbf{M}\|) = O_P(p^{2-2\delta}\ell_n) = O_P(p^{2-3\delta/2}n^{-1/2}),$$

which is Theorem 2(i). For $j = r+1, \cdots, (k_0+1)r$, using decomposition (8.4), we have

$$\begin{aligned} K_1 &= O_P((p^{1-\delta/2}n^{-1/2} + pn^{-1})^2) = O_P(p^{2-\delta}n^{-1} + p^2n^{-2}) = O_P(p^{2-\delta}n^{-1}), \\ |K_2| &= O_P(\|\widehat{\mathbf{M}} - \mathbf{M}\| \cdot \|\widehat{\mathbf{B}} - \mathbf{B}\|) = O_P(p^{2-2\delta}\ell_n^2) = O_P(p^{2-\delta}n^{-1}), \\ |K_3| &= O_P(\|\widehat{\mathbf{B}} - \mathbf{B}\|^2 \cdot \|\mathbf{M}\|) = O_P(p^{2-2\delta}\ell_n^2) = O_P(p^{2-\delta}n^{-1}). \end{aligned}$$

Hence $\widehat{\lambda}_j = O_P(p^{2-2\delta}\ell_n^2) = O_P(p^{2-\delta}n^{-1})$, which is Theorem 2(ii).

For part (iii), we define

$$\mathbf{W}_y(k_0) = (\mathbf{\Sigma}_y(1), \cdots, \mathbf{\Sigma}_y(k_0)), \quad \widehat{\mathbf{W}}_y(k_0) = (\widehat{\mathbf{\Sigma}}_y(1), \cdots, \widehat{\mathbf{\Sigma}}_y(k_0)),$$

so that $\mathbf{M} = \mathbf{W}_y(k_0)\mathbf{W}_y(k_0)'$ and $\widehat{\mathbf{M}} = \widehat{\mathbf{W}}_y(k_0)\widehat{\mathbf{W}}_y(k_0)'$. We define $\widehat{\mathbf{W}}_x(k_0), \widehat{\mathbf{W}}_{x\epsilon}(k_0)$, $\widehat{\mathbf{W}}_{\epsilon x}(k_0)$ and $\widehat{\mathbf{W}}_\epsilon(k_0)$ similarly. Then we can write

$$\widehat{\mathbf{W}}_y(k_0) = M_1 + M_2 + \widehat{\mathbf{W}}_\epsilon(k_0),$$

where $M_1 = \mathbf{A}(\widehat{\mathbf{W}}_x(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{A}') + \widehat{\mathbf{W}}_{x\epsilon}(k_0))$, $M_2 = \widehat{\mathbf{W}}_{\epsilon x}(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{A}')$. It is easy to see that

$$\text{rank}(M_1) \le r, \quad \text{rank}(M_2) \le k_0 r,$$

so that $\text{rank}(M_1 + M_2) \le (k_0+1)r$. This implies that

$$\sigma_j(M_1 + M_2) = 0, \quad \text{for} \quad j = (k_0+1)r+1, \cdots, p.$$

Then by Theorem 3.3.16(a) of Horn and Johnson (1991), for $j = (k_0+1)r+1, \cdots, p$,

$$\begin{aligned}\widehat{\lambda}_j = \lambda_j(\widehat{\mathbf{M}}) = \sigma_j^2(\widehat{\mathbf{W}}_y(k_0)) &\le (\sigma_j(M_1+M_2) + \sigma_1(\widehat{\mathbf{W}}_\epsilon(k_0)))^2 \\ &= \sigma_1^2(\widehat{\mathbf{W}}_\epsilon(k_0)) \\ &\le \sum_{k=1}^{k_0} \|\widehat{\mathbf{\Sigma}}_\epsilon(k)\|^2 \\ &= O_P(p^2 n^{-2}),\end{aligned}$$

where the last line follows from (8.5). This completes the proof of the theorem. □

To prove Theorem 3, 4 and Corollary 3, we need 4 lemmas. The first two are mathematical tools that we are using, and the third one presents the rates of λ_j for $j = 1, \cdots, r$. The following is Lemma 1 of Lam, Yao and Bathia (2010).

Lemma 2 *Suppose* $\mathbf{A}$ *and* $\mathbf{A}+\mathbf{E}$ *are* $n \times n$ *symmetric matrices and that*

$$\mathbf{Q} = [\mathbf{Q}_1 \;\; \mathbf{Q}_2] \quad (\mathbf{Q}_1 \;\; \textit{is} \;\; n \times r, \; \mathbf{Q}_2 \;\; \textit{is} \;\; n \times (n-r))$$

is an orthogonal matrix such that $\mathrm{span}(\mathbf{Q}_1)$ *is an invariant subspace for* $\mathbf{A}$ *(that is,* $\mathbf{A} \cdot \mathrm{span}(\mathbf{Q}_1) \subset \mathrm{span}(\mathbf{A})$*). Partition the matrices* $\mathbf{Q}'\mathbf{A}\mathbf{Q}$ *and* $\mathbf{Q}'\mathbf{E}\mathbf{Q}$ *as follows:*

$$\mathbf{Q}'\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix} \qquad \mathbf{Q}'\mathbf{E}\mathbf{Q} = \begin{pmatrix} \mathbf{E}_{11} & \mathbf{E}_{21}' \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix}.$$

If $\mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2) := \min_{\lambda \in \lambda(\mathbf{D}_1),\, \mu \in \lambda(\mathbf{D}_2)} |\lambda - \mu| > 0$*, where* $\lambda(M)$ *denotes the set of eigenvalues of the matrix* M*, and*

$$\|\mathbf{E}\| \le \frac{\mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2)}{5},$$

then there exists a matrix $\mathbf{P} \in \mathbb{R}^{(n-r)\times r}$ *with*

$$\|\mathbf{P}\| \le \frac{4}{\mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2)} \|\mathbf{E}_{21}\|$$

such that the columns of $\widehat{\mathbf{Q}}_1 = (\mathbf{Q}_1 + \mathbf{Q}_2\mathbf{P})(\mathbf{I} + \mathbf{P}'\mathbf{P})^{-1/2}$ *define an orthonormal basis for a subspace that is invariant for* $\mathbf{A}+\mathbf{E}$*.*

The following is the Wielandt's inequality.

Lemma 3 *Let* $\mathbf{A} \in \mathbb{R}^{r\times r}$ *be a symmetric matrix such that*

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}' & \mathbf{D} \end{pmatrix}, \quad \textit{with} \;\; \mathbf{B} \in \mathbb{R}^{r_1 \times r_1}, \; \mathbf{D} \in \mathbb{R}^{r_2 \times r_2} \;\; \textit{and} \;\; \lambda_{r_1}(\mathbf{B}) > \lambda_1(\mathbf{D}).$$

Then for $1 \le j \le r_2$*,*

$$0 \le \lambda_j(\mathbf{D}) - \lambda_{r_1+j}(\mathbf{A}) \le \frac{\lambda_1(\mathbf{C}\mathbf{C}')}{\lambda_{r_1}(\mathbf{B}) - \lambda_j(\mathbf{D})}.$$

Hereafter we suppress k_0 and write $\mathbf{W}_y = \mathbf{W}_y(k_0)$, $\mathbf{W}_i = \mathbf{W}_i(k_0)$, $\mathbf{W}_{21} = \mathbf{W}_{21}(k_0)$ and $\mathbf{W}_{12} = \mathbf{W}_{12}(k_0)$. Similarly $\mathbf{W}_{y^*} = \mathbf{W}_{y^*}(k_0)$.

Lemma 4 *For model (6.1) and* $\mathbf{M}$ *defined in (4.3), under conditions C1-C4,C5',C6', we have*

$$\lambda_j \asymp \begin{cases} p^2, & \text{for } j = 1, \cdots, r_1; \\ p^{2-2\delta_2}, & \text{if } \|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2}), \quad \text{for } j = r_1+1, \cdots, r; \\ p^{2-2c\delta_2}, & \text{if } \|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}, \quad 1/2 \le c \le 1, \quad \text{and} \\ & \|\mathbf{W}_1\mathbf{W}_{21}'\| \le q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\|, \ 0 \le q < 1, \text{ for } j = r_1+1, \cdots, r. \end{cases}$$

For model (6.1), and $\mathbf{M}^*$ *defined in section 6 by* $\mathbf{y}_t^*$ *in (6.2), we have*

$$\lambda_j^* \asymp p^{2-2\delta_2} \quad \text{for } j = 1, \cdots, r_2.$$

Proof of Lemma 4. With the notations in the proof of Theorem 2 and section 6, we have $\mathbf{M} = \mathbf{W}_y\mathbf{W}_y'$, where we can write, by assumption C6',

$$\begin{aligned} \mathbf{W}_y &= \mathbf{A}_1\mathbf{L}_1 + \mathbf{A}_2\mathbf{L}_2, \quad \text{with} \\ \mathbf{L}_1 &= \mathbf{W}_1(\mathbf{I}_{k_0} \otimes \mathbf{A}_1') + \mathbf{W}_{12}(\mathbf{I}_{k_0} \otimes \mathbf{A}_2'), \\ \mathbf{L}_2 &= \mathbf{W}_2(\mathbf{I}_{k_0} \otimes \mathbf{A}_2') + \mathbf{W}_{21}(\mathbf{I}_{k_0} \otimes \mathbf{A}_1'). \end{aligned}$$

To find the order of λ_j for $j = 1, \cdots, r_1$, using a standard singular value inequality, for $j = 1, \cdots, r_1$,

$$\lambda_j = \lambda_j(\mathbf{M}) = \sigma_j^2(\mathbf{W}_y) \ge \{\sigma_j(\mathbf{A}_1\mathbf{L}_1) - \sigma_1(\mathbf{A}_2\mathbf{L}_2)\}^2 = \{\sigma_j(\mathbf{L}_1) - \sigma_1(\mathbf{L}_2)\}^2,$$

where assumptions C5' ensures that $\mathbf{L}_1$ has the j-th singular value dominates those of $\mathbf{L}_2$, since $\delta_2 > \delta_1 = 0$. Hence, using the same singular value inequality,

$$\begin{aligned} \lambda_j \ge \{\sigma_j(\mathbf{L}_1) - \sigma_1(\mathbf{L}_2)\}^2 &\asymp \sigma_j^2(\mathbf{L}_1) \ge \{\sigma_j(\mathbf{W}_1(\mathbf{I}_{k_0} \otimes \mathbf{A}_1')) - \sigma_1(\mathbf{W}_{12}(\mathbf{I}_{k_0} \otimes \mathbf{A}_2'))\}^2 \\ &\asymp \sigma_j^2(\mathbf{W}_1(\mathbf{I}_{k_0} \otimes \mathbf{A}_1')) = \sigma_j^2(\mathbf{W}_1) = \lambda_j\left(\sum_{k=1}^{k_0} \mathbf{\Sigma}_1(k)\mathbf{\Sigma}_1(k)'\right) \\ &\ge \lambda_j(\mathbf{\Sigma}_1(1)\mathbf{\Sigma}_1(1)') \ge \|\mathbf{\Sigma}_1(1)\|_{\min} \asymp p^2, \end{aligned}$$

where the start of the second line follows from condition C5' and the fact that $\|\mathbf{W}_{12}\|, \|\mathbf{W}_{21}\| = O(p^{1-\delta_2/2}) = o(\|\mathbf{W}_1\|)$. This shows that $\lambda_j \asymp p^2$ for $j = 1, \cdots, r_1$, which is the largest possible order.

Also, $\mathbf{M} = \mathbf{W}_y\mathbf{W}_y' = \mathbf{A}\mathbf{L}\mathbf{A}'$, where

$$\mathbf{A} = (\mathbf{A}_1 \ \mathbf{A}_2), \quad \mathbf{L} = \begin{pmatrix} \mathbf{L}_1\mathbf{L}_1' & \mathbf{L}_1\mathbf{L}_2' \\ \mathbf{L}_2\mathbf{L}_1' & \mathbf{L}_2\mathbf{L}_2' \end{pmatrix}, \tag{8.6}$$

and $\mathbf{M}$ and $\mathbf{L}$ shares the same eigenvalues, as $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Hence $\lambda_j = \lambda_j(\mathbf{M}) = \lambda_j(\mathbf{L})$. By Lemma 3 applying on $\mathbf{L}$, we have for $j = 1, \cdots, r_2$,

$$\lambda_{r_1+j}(\mathbf{L}) \leq \lambda_j(\mathbf{L}_2\mathbf{L}_2') = \sigma_j^2(\mathbf{L}_2), \tag{8.7}$$

$$\lambda_{r_1+j}(\mathbf{L}) \geq \sigma_j^2(\mathbf{L}_2) - \frac{\sigma_1^2(\mathbf{L}_1\mathbf{L}_2')}{\sigma_{r_1}^2(\mathbf{L}_1) - \sigma_1^2(\mathbf{L}_2)}. \tag{8.8}$$

From the definition of $\mathbf{L}_1$ and $\mathbf{L}_2$, we have $\mathbf{L}_1\mathbf{L}_2' = \mathbf{W}_1\mathbf{W}_{21}' + \mathbf{W}_2\mathbf{W}_{12}'$. By condition C5' and the Cauchy-Schwarz inequality for spectral norm, we have

$$\|\mathbf{W}_2\mathbf{W}_{12}'\| \leq \|\mathbf{W}_2\|\|\mathbf{W}_{12}\| = O(p^{1-\delta_2} \cdot p^{1-\delta_2/2}) = O(p^{2-3\delta_2/2}) = o(\|\mathbf{L}_1\|\|\mathbf{L}_2\|),$$

since $\|\mathbf{L}_1\| = \sigma_1(\mathbf{L}_1) \asymp \sigma_1(\mathbf{W}_1) \asymp p$, and $\|\mathbf{L}_2\| = \sigma_1(\mathbf{L}_2) \asymp \max(\|\mathbf{W}_2\|, \|\mathbf{W}_{21}\|)$ with $\|\mathbf{W}_2\| \asymp p^{1-\delta_2}$. Hence asymptotically we can always find a constant $C_2 > 0$ such that

$$\|\mathbf{W}_2\mathbf{W}_{12}'\| \leq C_2\|\mathbf{L}_1\|\|\mathbf{L}_2\|. \tag{8.9}$$

Case 1. If $\|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}$ ($\asymp \|\mathbf{W}_{21}\|$ by condition C5') for $1/2 \leq c < 1$, then $\|\mathbf{W}_2\| \asymp p^{1-\delta_2} = o(\|\mathbf{W}_{21}\|)$. Hence $\|\mathbf{L}_2\| \asymp \|\mathbf{W}_{21}\| \asymp p^{1-c\delta_2}$. The assumption $\|\mathbf{W}_1\mathbf{W}_{21}'\| \leq q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\|$ for $0 < q < 1$ then ensures that

$$\begin{aligned}
\|\mathbf{W}_1\mathbf{W}_{21}'\| &\leq q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\| = (q\sigma_{r_1}(\mathbf{W}_1)/\sigma_1(\mathbf{W}_1))\|\mathbf{W}_1\|\|\mathbf{W}_{21}\| \\
&\leq (q\sigma_{r_1}(\mathbf{W}_1)/\sigma_1(\mathbf{W}_1))(\|\mathbf{L}_1\| + \|\mathbf{W}_{12}\|)(\|\mathbf{L}_2\| + \|\mathbf{W}_2\|) \\
&= (q\sigma_{r_1}(\mathbf{W}_1)/\sigma_1(\mathbf{W}_1))\|\mathbf{L}_1\|\|\mathbf{L}_2\|(1 + o(1))(1 + o(1)) \\
&\leq C_1\|\mathbf{L}_1\|\|\mathbf{L}_2\|,
\end{aligned} \tag{8.10}$$

where $0 < C_1 < \sigma_{r_1}(\mathbf{L}_1)/\sigma_1(\mathbf{L}_1)$, when n, p are sufficiently large. Combining (8.9) and (8.10), we then have

$$\begin{aligned}
\sigma_1(\mathbf{L}_1\mathbf{L}_2') &\leq \|\mathbf{W}_1\mathbf{W}_{21}'\| + \|\mathbf{W}_2\mathbf{W}_{12}'\| \leq (C_1 + C_2)\sigma_1(\mathbf{L}_1)\sigma_1(\mathbf{L}_2) \\
&= C\sigma_1(\mathbf{L}_1)\sigma_1(\mathbf{L}_2),
\end{aligned} \tag{8.11}$$

where C_2 in (8.9) is chosen so that $0 < C = C_1 + C_2 < \sigma_{r_1}(\mathbf{L}_1)/\sigma_1(\mathbf{L}_1)$.

Case 2. If $\|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2})$, then $\|\mathbf{W}_{21}\| = o(\|\mathbf{W}_2\|)$. Hence $\|\mathbf{L}_2\| \asymp \|\mathbf{W}_2\| \asymp p^{1-\delta_2}$. Hence

$$\|\mathbf{W}_1\mathbf{W}_{21}'\| \leq \|\mathbf{W}_1\|\|\mathbf{W}_{21}\| = o(\|\mathbf{L}_1\|\|\mathbf{L}_2\|),$$

and thus asymptotically we can always find a constant $C_1 > 0$ such that

$$\|\mathbf{W}_1\mathbf{W}_{21}'\| \leq C_1\|\mathbf{L}_1\|\|\mathbf{L}_2\|. \tag{8.12}$$

Combining (8.9) and (8.12), and choosing C_1 and C_2 to be small enough, (8.11) will hold.

With (8.11), (8.8) becomes

$$\lambda_{r_1+j}(\mathbf{L}) \geq \sigma_j^2(\mathbf{L}_2) - \frac{C^2\sigma_1^2(\mathbf{L}_1)\sigma_1^2(\mathbf{L}_2)}{\sigma_{r_1}^2(\mathbf{L}_1) - \sigma_1^2(\mathbf{L}_2)} = \sigma_j^2(\mathbf{L}_2)\left(1 - \frac{C^2}{\sigma_{r_1}^2(\mathbf{L}_1)/\sigma_1^2(\mathbf{L}_1) - o(1)}\right) \geq C'\sigma_j^2(\mathbf{L}_2)$$

for some constant $C' > 0$. Together with (8.7), we have $\lambda_{r_1+j}(\mathbf{L}) \asymp \sigma_j^2(\mathbf{L}_2)$ for $j = 1, \cdots, r_2$.

Hence, if $\|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2})$, then $\mathbf{W}_2$ dominates and from Case 2 above,

$$\lambda_{r_1+j}(\mathbf{L}) \asymp \sigma_j^2(\mathbf{L}_2) \asymp \sigma_j^2(\mathbf{W}_2) \asymp p^{2-2\delta_2}.$$

If $\|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}$ for $1/2 \leq c < 1$, then $\mathbf{W}_{21}$ dominates, and from Case 1 above,

$$\lambda_{r_1+j}(\mathbf{L}) \asymp \sigma_j^2(\mathbf{L}_2) \asymp \sigma_j^2(\mathbf{W}_{21}) \asymp p^{2-2c\delta_2},$$

which completes the proof for λ_j, $j = 1, \cdots, r$.

For λ_j^* with $j = 1, \cdots, r_2$, note that $\mathbf{M}^* = \mathbf{W}_{y^*}\mathbf{W}_{y^*}'$, where $\mathbf{W}_{y^*}$ is formed similar to $\mathbf{W}_y$, but by $\mathbf{y}_t^{0*} = \mathbf{y}_t - \mathbf{A}_1\mathbf{A}_1'\mathbf{y}_t$ (similar to (6.2), but $\widehat{\mathbf{A}}_1$ is replaced by $\mathbf{A}_1$). With condition C6', we can show by simple algebra that

$$\mathbf{W}_{y^*} = \mathbf{A}_2\mathbf{W}_2(\mathbf{I}_{k_0} \otimes \mathbf{A}_2'). \tag{8.13}$$

Hence,

$$\lambda_j^* = \sigma_j^2(\mathbf{W}_{y^*}) = \sigma_j^2(\mathbf{W}_2) \asymp p^{2-2\delta_2},$$

which completes the proof of the lemma. □

Proof of Theorem 3. We can write $\mathbf{M} = (\mathbf{A}_1\ \mathbf{A}_2\ \mathbf{B})\mathbf{D}(\mathbf{A}_1\ \mathbf{A}_2\ \mathbf{B})'$, where $\mathbf{B}$ is the orthogonal complement of $\mathbf{A} = (\mathbf{A}_1\ \mathbf{A}_2)$, and $\mathbf{D}$ is diagonal with $\mathbf{D} = \mathrm{diag}(\mathbf{D}_1,\ \mathbf{D}_2,\ \mathbf{0})$, where $\mathbf{D}_1$ contains λ_j for $j = 1, \cdots, r_1$ and $\mathbf{D}_2$ contains λ_j for $j = r_1 + 1, \cdots, r$. Hence by Lemma 4, using the $\mathrm{sep}(\cdot, \cdot)$ notation from Lemma 2,

$$\mathrm{sep}\Big(\mathbf{D}_1, \begin{pmatrix} \mathbf{D}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\Big) \asymp p^2, \tag{8.14}$$

$$\mathrm{sep}\Big(\mathbf{D}_2, \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\Big) \asymp \begin{cases} p^{2-2\delta_2}, & \text{if } \|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2}); \\ p^{2-2c\delta_2}, & \text{if } \|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2},\ 1/2 \leq c < 1, \text{ and} \\ & \|\mathbf{W}_1\mathbf{W}_{21}'\| \leq q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\|,\ 0 \leq q < 1. \end{cases} \tag{8.15}$$

Using Lemma 2, if we can show that for $i, j = 1, 2$ and $i \neq j$,

$$\|\mathbf{E}_{21}\| := \|\mathbf{A}_i'(\widehat{\mathbf{M}} - \mathbf{M})(\mathbf{A}_j\ \mathbf{B})\| = o_P\left(\mathrm{sep}\Big(\mathbf{D}_i, \begin{pmatrix} \mathbf{D}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\Big)\right), \tag{8.16}$$

then with similar arguments as in the proof of Theorem 1 in Lam, Yao and Bathia (2010), we can conclude that

$$\|\widehat{\mathbf{A}}_i - \mathbf{A}_i\| = O_P\left(\|\mathbf{A}_i'(\widehat{\mathbf{M}} - \mathbf{M})(\mathbf{A}_j\ \mathbf{B})\| / \mathrm{sep}\Big(\mathbf{D}_i, \begin{pmatrix} \mathbf{D}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\Big)\right). \tag{8.17}$$

Hence it remains to find the rate for $\|\mathbf{A}_i'(\widehat{\mathbf{M}}-\mathbf{M})(\mathbf{A}_j\ \mathbf{B})\|$. Since $\mathbf{MB}=\mathbf{0}$, we have $\|\mathbf{E}_{21}\| := \|\mathbf{A}_i'(\widehat{\mathbf{M}}-\mathbf{M})(\mathbf{A}_j\ \mathbf{B})\| = \|(\mathbf{A}_i'(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{A}_j\ \ \mathbf{A}_i'\widehat{\mathbf{M}}\mathbf{B})\| \le \|\mathbf{A}_i'(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{A}_j\| + \|\mathbf{A}_i'\widehat{\mathbf{M}}\mathbf{B}\|$.

Both terms can be decomposed into sum of more terms. When $i=1,j=2$, by some simple (but tedious) algebra, using condition C5' and C6'(i), we can show that $\|\mathbf{A}_1'\widehat{\mathbf{M}}\mathbf{B}\|$ is dominating, with the dominating term

$$\|\widehat{\mathbf{\Sigma}}_1(k)\widehat{\mathbf{\Sigma}}_{\epsilon 1}(k)'\mathbf{B}\| \le \|\widehat{\mathbf{\Sigma}}_1(k)\|\cdot\|\widehat{\mathbf{\Sigma}}_{\epsilon 1}(k)\| = O_P(p\cdot pn^{-1/2}) = O_P(p^2n^{-1/2}) = o(p^2),$$

Hence noting (8.14), criterion (8.16) is satisfied. Therefore using (8.17),

$$\|\widehat{\mathbf{A}}_1-\mathbf{A}_1\| = O_P(p^2n^{-1/2}/p^2) = O_P(n^{-1/2}).$$

For $i=2,j=1$, $\|\mathbf{A}_2'(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{A}_1\|$ is dominating, with the dominating term

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_{21}(k)\widehat{\mathbf{\Sigma}}_1(k)'-\mathbf{\Sigma}_{21}(k)\mathbf{\Sigma}_1(k)'\| &\le \|\widehat{\mathbf{\Sigma}}_{21}(k)(\widehat{\mathbf{\Sigma}}_1(k)-\mathbf{\Sigma}_1(k))\| + \|(\widehat{\mathbf{\Sigma}}_{21}(k)-\mathbf{\Sigma}_{21}(k))\mathbf{\Sigma}_1(k)\| \\ &= O_P((p^{1-\delta_2/2}n^{-1/2}+p^{1-c\delta_2})\cdot pn^{-1/2}+p^{1-\delta_2/2}n^{-1/2}\cdot p) \\ &= O_P(p^{2-\delta_2/2}n^{-1/2}) = o_P(p^{2-2c\delta_2}) \ \text{ for } 1/2\le c\le 1,\end{aligned}$$

by our assumption in the theorem that $\nu_n\to 0$, and $c=1$ represents the case $\|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2})$. In view of (8.15), criterion (8.16) is satisfied. Therefore using (8.17),

$$\|\widehat{\mathbf{A}}_2-\mathbf{A}_2\| = O_P(p^{2-\delta_2/2}n^{-1/2}/p^{2-2c\delta_2}) = O_P(p^{(2c-1/2)\delta_2}n^{-1/2}),$$

which is the rate given in the theorem. This completes the proof for the original procedure.

For the two-step procedure, the matrix $\mathbf{M}^* = (\mathbf{A}_2\ \mathbf{B}^*)\mathbf{D}^*(\mathbf{A}_2\ \mathbf{B}^*)'$, where $\mathbf{B}^*$ is the orthogonal complement of $\mathbf{A}_2$, and $\mathbf{D}^*$ is diagonal with $\mathbf{D}^* = \mathrm{diag}(\mathbf{D}_2^*,\mathbf{0})$. The matrix $\mathbf{D}_2^*$ contains λ_j^* for $j=1,\cdots,r_2$, so that by Lemma 4,

$$\mathrm{sep}(\mathbf{D}_2^*,\mathbf{0}) \asymp p^{2-2\delta_2}.$$

We can argue similar to the above (details omitted) to conclude that

$$\|\widetilde{\mathbf{A}}_2-\mathbf{A}_2\| = O_P(\|\mathbf{A}_2'(\widehat{\mathbf{M}}^*-\mathbf{M}^*)\mathbf{B}^*\|/\mathrm{sep}(\mathbf{D}_2^*,\mathbf{0})).$$

The dominating term in $\|\mathbf{A}_2'(\widehat{\mathbf{M}}^*-\mathbf{M}^*)\mathbf{B}^*\|$ can be shown to be, defining $\widehat{\mathbf{H}}_1 := \mathbf{I}-\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1'$,

$$\begin{aligned}\|\mathbf{A}_2'\widehat{\mathbf{H}}_1\mathbf{A}_2\widehat{\mathbf{\Sigma}}_2(k)\mathbf{A}_2'\widehat{\mathbf{H}}_1\mathbf{A}_2\widehat{\mathbf{\Sigma}}_{\epsilon 2}(k)'\widehat{\mathbf{H}}_1\mathbf{B}^*\| &\le \|\mathbf{\Sigma}_2(k)\|\cdot\|\mathbf{\Sigma}_{\epsilon 2}(k)\| \\ &= O_P(p^{1-\delta_2}\cdot p^{1-\delta_2/2}n^{-1/2}) = O_P(p^{2-3\delta_2/2}n^{-1/2}).\end{aligned}$$

Hence we have

$$\|\widetilde{\mathbf{A}}_2-\mathbf{A}_2\| = O_P(p^{2-3\delta_2/2}n^{-1/2}/p^{2-2\delta_2}) = O_P(p^{\delta_2/2}n^{-1/2}),$$

which completes the proof of the theorem. □

Proof of Theorem 4. Similar to (8.1), with conditions C3, C5' and C6', we can show that (details omitted) for model (6.1) and $k = 1, \cdots, k_0$,

$$\begin{aligned} \|\mathbf{\Sigma}_y\| &= O(\|\mathbf{\Sigma}_1(k)\|) = O(p), \\ \|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\| &= O_P(pn^{-1/2} + \|\widehat{\mathbf{\Sigma}}_\epsilon(k)\|) = O_P(pn^{-1/2}), \end{aligned}$$

where the last line follows from $\|\widehat{\mathbf{\Sigma}}_\epsilon(k)\| = O_P(pn^{-1})$ when $n = O(p)$ and conditions C7 and C8 are satisfied (see proof of Theorem 2 for details). Hence by (8.1),

$$\begin{aligned} \|\widehat{\mathbf{M}} - \mathbf{M}\| &= O_P(\|\mathbf{\Sigma}_y(k)\| \cdot \|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\|) = O_P(p^2 n^{-1/2}), \\ \|\mathbf{M}\| &= O(\|\mathbf{\Sigma}_y(k)\|^2) = O(p^2). \end{aligned} \tag{8.18}$$

By Theorem 3, using the notations in (8.3), we have

$$\|\widehat{\boldsymbol{\gamma}}_j - \mathbf{a}_j\| \le \begin{cases} \|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\| = O_P(n^{-1/2}), & \text{for } j = 1, \cdots, r_1; \\ \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(\nu_n), & \text{for } j = r_1 + 1, \cdots, r; \\ \|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(\nu_n), & \text{for } j = r + 1, \cdots, p, \end{cases} \tag{8.19}$$

where the rate for $\|\mathbf{B} - \mathbf{B}\|$ can be found similar to the proof of Theorem 3, and is thus omitted. With (8.18) and (8.19), the decomposition (8.3) is dominated by $\|I_3\|, \|I_4\|, \|I_5\| = O_P(p^2 n^{-1/2})$, so that $\|\widehat{\lambda}_j - \lambda_j\| = O_P(p^2 n^{-1/2})$. The rate $\lambda_j \asymp p^2$ is given by Lemma 4.

For $j = r_1 + 1, \cdots, r$, the same decomposition (8.3) together with (8.18) and (8.19) gives dominating terms $\|I_2\| = O_P(p^2 \nu_n^2)$ and $\|I_4\| = O_P(p^2 n^{-1/2})$, so that $\|\widehat{\lambda}_j - \lambda_j\| = O_P(p^2(n^{-1/2} + \nu_n^2))$. The rate for λ_j is given by Lemma 4.

For $j = r+1, \cdots, p$, the decomposition (8.4) together with (8.18) and (8.19) gives dominating term $\|K_3\| = O_P(\|\widehat{\mathbf{B}} - \mathbf{B}\| \cdot \|\mathbf{M}\|) = O_P(p^2 \nu_n^2)$. Hence $\widehat{\lambda}_j = O_P(p^2 \nu_n^2)$. This completes the proof for the original procedure.

For the two-step procedure, the decomposition (8.3) and (8.4) are both valid, with $\widehat{\mathbf{M}}$ and $\mathbf{M}$ replaced by $\widehat{\mathbf{M}}^*$ and $\mathbf{M}^*$ respectively, and $\widehat{\boldsymbol{\gamma}}_j$ replaced by $\widehat{\boldsymbol{\gamma}}_j^*$, the unit eigenvector of $\widehat{\mathbf{M}}^*$ for the j-th largest eigenvalue λ_j^* of $\mathbf{M}^*$. Since $\mathbf{M}^* = \mathbf{W}_{y^*}\mathbf{W}_{y^*}'$ and $\widehat{\mathbf{\Sigma}}_{y^*}(k) = \widehat{\mathbf{H}}_1 \widehat{\mathbf{\Sigma}}_y(k) \widehat{\mathbf{H}}_1$, using (8.13), we can show that (details omitted)

$$\begin{aligned} \|\mathbf{M}^*\| &= O(p^{2-2\delta_2}), \\ \|\widehat{\mathbf{M}}^* - \mathbf{M}^*\| &= O_P(\|\widehat{\mathbf{\Sigma}}_{y^*}(k) - \mathbf{\Sigma}_{y^*}(k)\| \cdot \|\mathbf{\Sigma}_{y^*}(k)\|) \\ &= O_P(p^{2-3\delta_2/2} n^{-1/2}) = O_P(p^{2-2\delta_2}\kappa_n). \end{aligned} \tag{8.20}$$

Let $\mathbf{a}_j^*$ be such that $\mathbf{A}_2 = (\mathbf{a}_1^*, \cdots, \mathbf{a}_{r_2}^*)$, $\mathbf{B}^* = (\mathbf{a}_{r_2+1}^*, \cdots, \mathbf{a}_p^*)$. Then we can show that

$$\|\widehat{\boldsymbol{\gamma}}_j^* - \mathbf{a}_j^*\| \le \begin{cases} \|\widetilde{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(\kappa_n), & \text{if } j = 1, \cdots, r_2; \\ \|\widetilde{\mathbf{B}}^* - \mathbf{B}^*\| = O_P(\kappa_n), & \text{if } j = r_2 + 1, \cdots, p. \end{cases} \tag{8.21}$$

For $j = 1, \cdots, r_2$, with (8.20) and (8.21), the dominating terms for the decomposition (8.3) are $\|I_3\|, \|I_4\|, \|I_5\| = O_P(p^{2-2\delta_2}\kappa_n)$, so that $\|\widehat{\lambda}_j^* - \lambda_j^*\| = O_P(p^{2-2\delta_2}\kappa_n)$. The rate for λ_j^* is given in Lemma 4.

For $j = r_2 + 1, \cdots, p$, with (8.20) and (8.21), all terms of the decomposition (8.4) have the same rate $O_P(p^{2-2\delta_2}\kappa_n^2)$, so that $\widehat{\lambda}_j^* = O_P(p^{2-2\delta_2}\kappa_n^2)$.

The last claim of the theorem is proved exactly the same as part (iii) of Theorem 2, and is omitted. □

Corollary 3 follows from Lemma 5 immediately.

Lemma 5 *Let conditions C1-C4, C5', C6', C7 and C8 hold. Then as $n, p \to \infty$ with $n = O(p)$, with ν_n and $\kappa_n \to 0$ the same as in Theorem 3, it holds that*

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \begin{cases} \asymp 1, & j = 1, \cdots, r_1 - 1; \\ = O_P(n^{-1/2} + \nu_n^2 + p^{-2\delta_2}), & j = r_1, \text{ if } \|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2}); \\ = O_P(n^{-1/2} + \nu_n^2 + p^{-2c\delta_2}), & \text{if } \|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2} \text{ for } 1/2 \le c < 1, \text{ and} \\ & \|\mathbf{W}_1\mathbf{W}_{21}'\| \le q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\| \text{ for } 0 \le q < 1. \end{cases}$$

Furthermore, if $\|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2})$ and $p^{5\delta_2/2}n^{-1/2} \to 0$, we have

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \begin{cases} \asymp 1, & j = r_1 + 1, \cdots, r - 1; \\ = O_P(p^{2\delta_2}\nu_n^2), & j = r. \end{cases}$$

If $\|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}$ for $1/2 \le c < 1$, $\|\mathbf{W}_1\mathbf{W}_{21}'\| \le q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\|$ for $0 \le q < 1$, and $p^{(3c-1/2)\delta_2}n^{-1/2} \to 0$, we have

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \begin{cases} \asymp 1, & j = r_1 + 1, \cdots, r - 1; \\ = O_P(p^{2c\delta_2}\nu_n^2), & j = r. \end{cases}$$

For the two-step procedure, let conditions C1-C4, C5', C6', C7 and C8 hold and $n = O(p)$. Then we have

$$\widehat{\lambda}_{j+1}^*/\widehat{\lambda}_j^* \begin{cases} \asymp 1, & j = 1, \cdots, r_2 - 1; \\ = O_P(\kappa_n^2), & j = r_2. \end{cases}$$

Proof of Lemma 5. We only need to find the asymptotic rate for each $\widehat{\lambda}_j$ and $\widehat{\lambda}_j^*$. The rate of each ratio can then be obtained from the results of Theorem 4.

For $j = 1, \cdots, r_1$, from Theorem 4, $\|\widehat{\lambda}_j - \lambda_j\| = O_P(p^2 n^{-1/2}) = o_P(\lambda_j)$, and hence $\widehat{\lambda}_j \asymp \lambda_j \asymp p^2$. Consider the case $\|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}$. For $j = r_1 + 1, \cdots, r$, since $|\widehat{\lambda}_j - \lambda_j| = O_P(p^2(n^{-1/2} + \nu_n^2))$, we have $\widehat{\lambda}_j \le \lambda_j + O_P(p^2(n^{-1/2} + \nu_n^2)) = O_P(p^{2-2c\delta_2} + p^2\nu_n^2 + p^2 n^{-1/2})$, and hence

$$\widehat{\lambda}_{r_1+1}/\widehat{\lambda}_{r_1} = O_P((p^{2-2c\delta_2} + p^2\nu_n^2 + p^2 n^{-1/2})/p^2) = O_P(n^{-1/2} + \nu_n^2 + p^{-2c\delta_2}).$$

The other case is proved similarly.

For $j = r_1 + 1, \cdots, r$, to make sure $\widehat{\lambda}_j$ will not be zero or close to zero, we need

$$|\widehat{\lambda}_j - \lambda_j| = O_P(p^2(n^{-1/2} + \nu_n^2)) = o_P(\lambda_j),$$

where $\lambda_j \asymp p^{2-2c\delta_2}$. Hence we need $p^2(n^{-1/2}+\nu_n^2) = o(p^{2-2c\delta_2})$, which is equivalent to the condition $p^{(3c-1/2)\delta_2}n^{-1/2} \to 0$. This implies $\widehat{\lambda}_j \asymp \lambda_j \asymp p^{2-2c\delta_2}$ for $j = r_1+1, \cdots, r$. Since $\widehat{\lambda}_j = O_P(p^2\nu_n^2)$ for $j = r+1, \cdots, p$, we then have

$$\widehat{\lambda}_{r+1}/\widehat{\lambda}_r = O_P(p^2\nu_n^2/p^{2-2c\delta_2}) = O_P(p^{2c\delta_2}\nu_n^2).$$

All other rates can be proved similarly, and thus are omitted. □

References

Anderson, T.W. (1963). The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika*, **28**, 1-25.

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

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

Bai, J. and Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, **25**, 52-60.

Bathia, N., Yao, Q. and Zieglemann, F. (2010). Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, **38**, 3352-3386.

Brillinger, D.R. (1981). *Time Series Analysis: Data Analysis and Theory* (2nd ed.). Holt, Rinehart & Winston, New York.

Chamberlain, G. (1983). Funds, factors, and diversification in arbitrage pricing models. *Econometrica*, **51**, 13051323.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, **51**, 1281-1304.

Chib, S. and Ergashev, B. (2009). Analysis of multifactor affine yield curve models. *Journal of the American Statistical Association*, **104**, 1324-1337.

Chudik, A. and Pesaran, M.H. (2009). Infinite dimensional VARs and factor models. Available at *http://ssrn.com/abstract=1079063*.

Deistler, M., Anderson, B., Chen, W. and Filler, A. (2009). Generalized linear dynamic factor models – an approach via singular autoregressions. *European Journal of Control*. (Invited submission.)

Dümbgen, L. (1995). A simple proof and refinement of Wielandt's eigenvalue inequality. *Statistics and Probability Letters*, **25**, 113-115.

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**, 540-554.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2004). The generalized dynamic-factor model: consistency and rates. *Journal of Econometrics* , **119**, 231-255.

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.

Geweke, J. (1977). The dynamic factor analysis of economic time series. In *Latent Variables in Socio-Economic Models*, ed. Aigner, D.J. and Goldberger A.S., Chapter 19, Amsterdam, North-Holland.

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

Hannan, E.J. (1970). *Multiple Time Series*. Wiley, New York.

Horn, R.A. and Johnson, C.R. (1991). *Topics in Matrix Analysis*, Cambridge University Press.

Lam, C., Yao, Q. and Bathia, N. (2010). Estimation for latent factor models for high-dimensional time series. *Biometrika*, *to appear*.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis* (2nd edition). Springer, Berlin.

Pan, J., Peña, D., Polonik, W. and Yao, Q. (2008). Modelling multivariate volatilities via common factors. Available at *http://stats.lse.ac.uk/q.yao/qyao.links/paper/pppy.pdf*.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, **95**, 356-379.

Pesaran, M.H. and Zaffaroni, P. (2008). Optimal asset allocation with factor models for large portfolios. Available at *http://ssrn.com/abstract=1145183*.

Péché, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probab. Theory Relat. Fields*, **143**, 481-516.

Peña, D. and Box, E.P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, **82**, 836-843.

Peña, D. and Poncela, P. (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, **136**, 1237-1257.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Academic Press, New York.

Priestley, M.B., Rao, T.S. and Tong, H. (1974). Applications of principal component analysis and factor analysis in the identification of multivariable systems. *IEEE Trans. Automat. Control*, **19**, 703-704.

Quah,D. and Sargent,T.J. (1993). A dynamic index model for large cross sections. Ch.7 in J.H. Stock and M.W. Walton (Eds.) *Business Cycles, Indicators and Forecasting*, NBER, 285-309.

Reinsel, G.C. (1997). *Elements of Mutivariate Time Series Analysis*. Springer, New York.

Ross, S. (1976). The arbitrage theory of capital asset pricing. *J. Fianance*, **13**, 341-360.

Sargent, T. J. and Sims, C.A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. In *New methods in business cycle research*, ed. Sims, C. et al, Minneapolis, Federal Reserve Bank of Minneapolis, 45-108.

Shapiro, D.E. and Switzer, P. (1989). Extracting time trends from mulpitle monitoring sites. Technical Report No.132, Department of Statistics, Stanford University.

Stock, J.H. and Watson, M.W. (1998). Diffusion indexed. *NBER Working Paper 6702*.

Stock, J.H. and Watson, M.W. (2002). Macroeconomic forecasting using diffusion indices. *Journal of Business & Economic Statistics*, **20**, 147-162.

Switzer, P. and Green, A.A. (1984). Min/Max autocorrelation factors for multivariate spatial imagery. Technical Report No.6, Department of Statistics, Stanford University.

Tiao, G.C. and Tsay, R.S. (1989). Model specification in multivariate time series (with discussion). *Journal of the Royal Statistical Society*, **B**, **51**, 157-213.

Wang, H. (2010). Factor Profiling for Ultra High Dimensional Variable Selection. Available at http://ssrn.com/abstract=1613452

Tao, M., Wang, Y., Yao, Q. and Zou, J. (2010). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *Journal of the American Statistical Association*, *to appear*.