Structural-Factor Modeling of High-Dimensional Time Series: Another Look at Approximate Factor Models with Diverging Eigenvalues

Zhaoxing Gao and Ruey S Tsay
Booth School of Business, University of Chicago

August 23, 2018

Abstract

This article proposes a new approach to modeling high-dimensional time series data by providing a simple and natural way to understand the mechanism of factor models. We treat a p-dimensional time series as a nonsingular linear transformation of certain common factors and structured idiosyncratic components. Unlike the approximate factor models, we allow the largest eigenvalues of the covariance matrix of the idiosyncratic components to diverge as the dimension p increases, which is reasonable in the high-dimensional setting. A white noise testing procedure for high-dimensional random vectors is proposed to determine the number of common factors under the assumption that the idiosyncratic term is a vector white noise. We also introduce a projected Principal Component Analysis (PCA) to eliminate the diverging effect of the noises. Asymptotic properties of the proposed method are established for both fixed p and diverging p as the sample size n tends to infinity. Both simulated and real examples are used to assess the performance of the proposed method. We also compare our method with two commonly used methods in the literature and find that the proposed approach not only provides interpretable results, but also performs well in out-of-sample forecasting.

Keywords: High dimension, Structured factor model, Eigen-analysis, Projected principal component analysis, Diverging eigenvalues, White noise test.

1 Introduction

Advances in information technology make large data sets widely accessible. In many applications, the data consist naturally of high-dimensional time series. For example, the returns of a large number of assets form a high-dimensional time series and play an important role in asset pricing, portfolio allocation, and risk management. Large panel time series data are commonplace in economics and biological studies. Environmental studies often employ high-dimensional time series consisting of a large number of pollution indexes collected at many monitoring stations and over periods of time. However, modeling high-dimensional time series is always challenging because the commonly used Vector Autoregressive (VAR) or Vector-Autoregressive Moving-Average (VARMA) models are not practically applicable when the dimension is high. In particular, unregularized VARMA models often suffer the difficulties of over-parameterization and lack of identifiability as discussed in Tiao and Tsay (1989), Lütkepohl (2006) and Tsay (2014). Therefore, dimension reduction or structural specification becomes a necessity in applications of high-dimensional time series. Indeed, various methods have been developed in the literature for multivariate time series analysis, including the scalar component models of Tiao and Tsay (1989), the LASSO regularization in VAR models by Shojaie and Michailidis (2010) and Song and Bickel (2011), the sparse VAR model based on partial spectral coherence by Davis et al. (2012), the factor modeling in Stock and Watson (2005), Bai and Ng (2002), Forni et al. (2005), Lam et al. (2011) and Lam and Yao (2012), among others. However, the complexity of the dynamical dependence in high-dimensional time series requires further investigation.

This paper marks a further development in factor modeling of high-dimensional time series. Factor models are commonly used in finance, economics, and statistics. For example, asset returns are often modeled as functions of a small number of factors, see Stock and Watson (1989) and Stock and Watson (1998). Macroeconomic variables of multiple countries are often found to have common movements, see Gregory and Head (1999) and Forni et al. (2000). From the statistical perspective, a modeling approach that can reveal the common structure of the series and provide accurate estimation of a specified model is highly valuable in understanding the dynamic relationships of the data. To the best of our knowledge, there are two main statistical procedures to estimate the common factors and the associated loading matrix. The first procedure is based on the principal component analysis (PCA), see Bai and Ng (2002) and Bai (2003). The other procedure is based on the eigen-analysis of the auto-covariance matrices, see Lam et al. (2011) and Lam and Yao (2012), among others. The

resulting model forms of the two procedures are similar, but the specified structure is slightly different. We briefly introduce the two procedures below.

Let $\mathbf{y}_t = (y_{1,t}, ..., y_{p,t})'$ be a p-dimensional zero-mean time series, where $\mathbf{A}'$ denotes the transpose of the vector or matrix $\mathbf{A}$. The approximate factor model for $\mathbf{y}_t$ assumes the form

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{1.1}$$

where $\mathbf{x}_t$ is a r-dimensional latent factor process, $\mathbf{A} \in R^{p\times r}$ is an associated factor loading matrix, $\boldsymbol{\varepsilon}_t$ is the idiosyncratic component, and $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_t$ are independent. In the econometric literature, $\boldsymbol{\varepsilon}_t$ is not necessarily a white noise series and the model is only asymptotically identifiable. See, for instance, Bai and Ng (2002) and Forni et al. (2005). On the other hand, statisticians often require $\boldsymbol{\varepsilon}_t$ being serially uncorrelated. See, for instance, Lam et al. (2011). Let $\widehat{\boldsymbol{\Sigma}}_y$ be the sample covariance matrix of $\mathbf{y}_t$, then $\widehat{\boldsymbol{\Sigma}}_y = \widehat{\mathbf{P}}\widehat{\mathbf{D}}\widehat{\mathbf{P}}'$, where $\widehat{\mathbf{P}}$ is an orthonormal matrix and $\widehat{\mathbf{D}}$ is a diagonal matrix with decreasing eigenvalues. The PCA estimators of $\mathbf{A}$ and $\mathbf{x}_t$ in Bai and Ng (2002) are denoted respectively as $\widehat{\mathbf{P}}_r\widehat{\mathbf{D}}_r^{1/2}$ and $\widehat{\mathbf{D}}_r^{-1/2}\widehat{\mathbf{P}}_r'\mathbf{y}_t$, where $\widehat{\mathbf{P}}_r$ consists of the first r columns of $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{D}}_r$ contains the corresponding r largest eigenvalues in $\widehat{\mathbf{D}}$. The number of factors r is determined by some information criterion. For details, see Bai and Ng (2002) and Bai (2003). In the time series literature, Lam et al. (2011) proposed a different approach. Let $\widehat{\boldsymbol{\Gamma}}_k$ be the sample auto-covariance matrix between $\mathbf{y}_t$ and $\mathbf{y}_{t-k}$ and $\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\boldsymbol{\Gamma}}_k\widehat{\boldsymbol{\Gamma}}_k'$ for some fixed positive integer k_0. Under the assumption that $\boldsymbol{\varepsilon}_t$ is a vector white noise, the authors estimate $\mathbf{A}$ by $\widehat{\mathbf{A}}$, which contains the eigenvectors of $\widehat{\mathbf{M}}$ associated with the r largest eigenvalues. The estimated common factors are $\widehat{\mathbf{x}}_t = \widehat{\mathbf{A}}'\mathbf{y}_t$ and the associated noises are $\widehat{\boldsymbol{\varepsilon}}_t = (\mathbf{I}_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}')\mathbf{y}_t$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. The number of common factors r is estimated by a method based on the ratios of the eigenvalues of $\widehat{\mathbf{M}}$. Asymptotic properties of the estimators of the two procedures have been derived by the proponents under certain regularity conditions. However, some fundamental issues remain unsolved:

- The PCA method may fail if the signal-to-noise ratio is low, which occurs often in applications. Consider, for instance, analysis of high-dimensional financial series. As the market and economic information accumulates, the noise is often increasing faster than the signal. See, for example, Black (1986).

- The estimated factor process $\widehat{\mathbf{x}}_t$ in Lam et al. (2011) includes the noise components. When the largest eigenvalues of the noise covariance are diverging, the resulting esti-

mators would deteriorate.

- The information criterion in Bai and Ng (2002) and the ratio-based method in Lam et al. (2011) may also fail if the largest eigenvalues of the covariance matrix of the noise are diverging. See the illustrations of the Assumptions on the approximate factor models in Bai and Ng (2002) (pp. 197).

- The sample covariance matrix of the estimated noises is singular if $r > 0$. This can easily be seen as the right side of Equation (1.1) consists of $r + p$ innovations whereas the data $\mathbf{y}_t$ is p-dimensional. Specifically, the common factors $\mathbf{f}_t$ have r innovations and the noise term has p innovations.

The first goal of this paper is to address the aforementioned issues from a different perspective. We propose a new factor model under which the observed high-dimensional time series $\mathbf{y}_t$ is a nonsingular linear transformation of a r-dimensional common factor process, which is dynamically dependent, and a $(p-r)$-dimensional idiosyncratic component, which is a white noise series. The new factor model is in line with that of Tiao and Tsay (1989) and Gao and Tsay (2018) and assumes that the idiosyncratic component is white noise in the sense that $\mathbf{y}_t$ has $(p-r)$ scalar components of order $(0,0)$. See Tiao and Tsay (1989) and Section 2 for details. However, the proposed modeling approach of this paper is different from those of Tiao and Tsay (1989)) and Gao and Tsay (2018), which employ canonical correlation analysis (CCA). To use CCA, Tiao and Tsay (1989) assumes p is fixed and Gao and Tsay (2018) considers $p = o(n^{1/2})$, where n is the sample size. On the other hand, we do not employ CCA in this paper and, hence, relax the constraints on p as n increases.

Similar to that of Lam et al. (2011) and Lam and Yao (2012), we first apply the eigenanalysis on certain auto-covariance matrices to obtain the loading matrix associated with common factors. But we propose a projected PCA method to estimate the loading matrix associated with the idiosyncratic component; see Section 2 for details. In addition, we propose a new method to estimate the common factors so that the resulting estimated common factors are not affected by the idiosyncratic component $\boldsymbol{\varepsilon}_t$. Specifically, in the presence of diverging noise components, we project the observed data into the orthogonal direction of those diverging noise components to mitigate the effect of the idiosyncratic component in estimating the common factors. Furthermore, to overcome the difficulties associated with the behavior of eigenvalues of a large random matrix, we consider a white noise testing procedure to determine the number of common factors. This testing procedure is found to

be more reliable than the information criterion and the ratio-based method currently used in the literature. Details of the testing procedure is given in Section 2.3.

Simulation studies are used to assess the performance of the proposed modeling procedure in finite samples and to compare it with the methods in Bai and Ng (2002) and Lam et al. (2011). The results show that the proposed method works well whereas the latter methods may encounter prominent estimation errors if the largest eigenvalues of the covariance matrix of the idiosyncratic component are diverging. We further apply the proposed method to two real data examples, and the numerical results suggest that the factors recovered by our approach not only have reasonable interpretations but also fare well in predictions. On the other hand, the factors recovered by the PCA method of Bai and Ng (2002) may contain white noise since PCA only deals with the covariance matrix without considering the dynamic dependence of the data, and the ratio-based method of Lam et al. (2011) often finds a single factor in practice, because the largest eigenvalue of $\widehat{\mathbf{M}}$ tends to be extremely large in the high dimensional case.

The contributions of this paper are multi-fold. First, the proposed new model is more flexible compared with the existing ones. In fact, it allows a variety of structures for the common factors and the idiosyncratic term. Second, the proposed estimation method can eliminate the effect of the idiosyncratic term in estimating the common factors. This is achieved by using the projected PCA method if the dimension p is low. When the dimension is high, we assume that a few largest eigenvalues of the covariance matrix of the idiosyncratic term are diverging, which is a reasonable assumption in the high-dimensional setting. The projected PCA then helps to mitigate the effect of the diverging part of the noise covariance matrix. Third, we propose a procedure based on a white noise test for multiple time series to determine the number of common factors r. Under the assumption that the idiosyncratic term is a vector white noise, the limiting distribution of the test statistic used is available in close form. This testing procedure is shown to be more reliable than the information criterion and the ratio-based method available in the literature.

The rest of the paper is organized as follows. We introduce the proposed model and estimation methodology in Section 2. In Section 3, we study the theoretical properties of the proposed model and its associated estimates. Numerical illustrations with both simulated and real data sets are reported in Section 4. Section 5 provides some discussions and concluding remarks. All technical proofs and an additional real example are relegated to an online supplement. Throughout the article, we use the following notation. $||\mathbf{u}||_2 = (\sum_{i=1}^{p} u_i^2)^{1/2}$

is the Euclidean norm of a p-dimensional vector $\mathbf{u} = (u_1, ..., u_p)'$ and $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. For a matrix $\mathbf{H} = (h_{ij})$, $|\mathbf{H}|_\infty = \max_{i,j} |h_{ij}|$, $\|\mathbf{H}\|_2 = \sqrt{\lambda_{\max}(\mathbf{H}'\mathbf{H})}$ is the operator norm, where $\lambda_{\max}(\cdot)$ denotes for the largest eigenvalue of a matrix, and $\|\mathbf{H}\|_{\min}$ is the square root of the minimum non-zero eigenvalue of $\mathbf{H}'\mathbf{H}$. The superscript $'$ denotes the transpose of a vector or matrix. Finally, we use the notation $a \asymp b$ to denote $a = O(b)$ and $b = O(a)$.

2 The Model and Methodology

2.1 Setting

Let $\mathbf{y}_t = (y_{1t}, ..., y_{pt})'$ be a p-dimensional time series. We assume $\mathbf{y}_t$ is observable with $E(\mathbf{y}_t) = 0$ and admits a latent structure:

$$\mathbf{y}_t = \mathbf{L} \begin{bmatrix} \mathbf{f}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix} = [\mathbf{L}_1, \mathbf{L}_2] \begin{bmatrix} \mathbf{f}_t \\ \boldsymbol{\varepsilon}_t \end{bmatrix} = \mathbf{L}_1 \mathbf{f}_t + \mathbf{L}_2 \boldsymbol{\varepsilon}_t, \tag{2.1}$$

where $\mathbf{L} \in R^{p \times p}$ is a full rank loading matrix, $\mathbf{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a r-dimensional factor process, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{vt})'$ is a v-dimensional white noise vector, and $r + v = p$. For meaningful dimension reduction, we assume r is a small fixed nonnegative integer. In addition, we also assume $\mathrm{Cov}(\mathbf{f}_t) = \mathbf{I}_r$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_t) = \mathbf{I}_v$, $\mathrm{Cov}(\mathbf{f}_t, \boldsymbol{\varepsilon}_t) = 0$, and no linear combination of $\mathbf{f}_t$ is serially uncorrelated.

The decomposition of model (2.1) is general in the sense that any finite-order VARMA time series $\mathbf{y}_t$ can always be written in Equation (2.1) via canonical correlation analysis between two constructed random vectors of $\mathbf{y}_t$ and its lagged variables. Details can be found in Tiao and Tsay (1989) and Gao and Tsay (2018). Under Equation (2.1), the dynamic dependence of $\mathbf{y}_t$ is driven by $\mathbf{f}_t$ if $r > 0$. Thus, $\mathbf{f}_t$ indeed consists of the common factors of $\mathbf{y}_t$. In the terminology of Tiao and Tsay (1989), (a) $\boldsymbol{\varepsilon}_t$ is a v-dimensional scalar component process of order (0,0) if $v > 0$, that is, $\mathrm{Cov}(\boldsymbol{\varepsilon}_t, \mathbf{y}_{t-j}) = \mathbf{0}$ for $j > 0$, and (b) no linear combination of $\mathbf{f}_t$ is a scalar component of order (0,0) if $r > 0$. Condition (b) is trivial because existence of any such linear combinations implies r can be reduced. Readers are referred to Tiao and Tsay (1989) for a formal definition of a scalar component of order (0,0). Condition (a) is equivalent to $\boldsymbol{\varepsilon}_t$ being a white noise under the traditional factor models, where $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_t$ are assumed to be independent.

We mention that Model (2.1) has been studied by Gao and Tsay (2018) if $\mathbf{y}_t$ follows a structural model consisting of trend, seasonal component, and irregular series. The irregular component of $\mathbf{y}_t$ is modeled by (2.1) using CCA. However, the method of CCA only works when $p < n$ or $p = o(n^{1/2})$, where n is the sample size. This restricts the applicability of the model . We relax such restrictions in this paper.

To study Model (2.1) in a more general setting and to provide sufficient statistical insights on the factor models, let $\mathbf{L}_1 = \mathbf{A}_1\mathbf{Q}_1$ and $\mathbf{L}_2 = \mathbf{A}_2\mathbf{Q}_2$, where $\mathbf{A}_1$ and $\mathbf{A}_2$ are two half orthonormal matrices, i.e. $\mathbf{A}_1'\mathbf{A}_1 = \mathbf{I}_r$ and $\mathbf{A}_2'\mathbf{A}_2 = \mathbf{I}_v$. This can be done via QR decomposition or singular value decomposition. Furthermore, let $\mathbf{x}_t = \mathbf{Q}_1\mathbf{f}_t$ and $\mathbf{e}_t = \mathbf{Q}_2\boldsymbol{\varepsilon}_t$, then Model (2.1) can be written as

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{x}_t + \mathbf{A}_2\mathbf{e}_t, \tag{2.2}$$

which is close to the traditional factor model in Equation (1.1). Some remarks are in order. First, even though $\mathbf{L}$ is of full rank, $\mathbf{A}_1$ is not orthogonal to $\mathbf{A}_2$ in general. Second, $\mathbf{A}_1$ and $\mathbf{x}_t$ are still not uniquely identified because we can replace $(\mathbf{A}_1, \mathbf{x}_t)$ by $(\mathbf{A}_1\mathbf{H}, \mathbf{H}'\mathbf{x}_t)$ for any orthonormal matrix $\mathbf{H} \in R^{r\times r}$. The same issue applies to $\mathbf{A}_2$ and $\mathbf{e}_t$. Nevertheless the linear space spanned by the columns of $\mathbf{A}_1$, denoted by $\mathcal{M}(\mathbf{A}_1)$, is uniquely defined. $\mathcal{M}(\mathbf{A}_1)$ is called the factor loading space. The linear space $\mathcal{M}(\mathbf{A}_2)$ can be defined similarly for the idiosyncratic component.

2.2 Estimation

To begin, we provide some rationale for the proposed estimation method. Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be the orthonormal complement of $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, i.e. $\mathbf{B}_1 \in R^{p\times v}$ and $\mathbf{B}_2 \in R^{p\times r}$ are two half orthonormal matrices satisfying $\mathbf{B}_1'\mathbf{A}_1 = \mathbf{0}$ and $\mathbf{B}_2'\mathbf{A}_2 = \mathbf{0}$. Denote $(\mathbf{A}_1, \mathbf{B}_1) = (\mathbf{a}_1, ..., \mathbf{a}_r, \mathbf{b}_1, ..., \mathbf{b}_v)$ and $(\mathbf{A}_2, \mathbf{B}_2) = (\mathbf{a}_{r+1}, ..., \mathbf{a}_p, \mathbf{b}_{v+1}, ..., \mathbf{b}_p)$, which are $p \times p$ matrices. It follows from Model (2.2) that

$$\mathbf{B}_1'\mathbf{y}_t = \mathbf{B}_1'\mathbf{A}_2\mathbf{e}_t, \tag{2.3}$$

and, hence, $\mathbf{B}_1'\mathbf{y}_t$ is a v-dimensional white noise process. Thus, for any column $\mathbf{b}_j$ of $\mathbf{B}_1$ with $1 \le j \le v$, $\{\mathbf{b}_j'\mathbf{y}_t, t = 0, \pm1, \ldots\}$ is a white noise process.

Unlike the traditional factor models, which assume $\mathbf{x}_t$ and $\mathbf{e}_s$ are uncorrelated for any t and s, we only require $\mathrm{Cov}(\mathbf{x}_t, \mathbf{e}_{t+j}) = 0$ for $j \ge \mathbf{0}$ in this paper. For $k \ge 0$, let

$$\boldsymbol{\Sigma}_y(k) = \mathrm{Cov}(\mathbf{y}_t, \mathbf{y}_{t-k}),\ \boldsymbol{\Sigma}_x(k) = \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_{t-k}),\ \boldsymbol{\Sigma}_{xe}(k) = \mathrm{Cov}(\mathbf{x}_t, \mathbf{e}_{t-k}),$$

be the covariance matrices of interest. It follows from (2.2) that

$$\mathbf{\Sigma}_y(k) = \mathbf{A}_1\mathbf{\Sigma}_x(k)\mathbf{A}_1' + \mathbf{A}_1\mathbf{\Sigma}_{xe}(k)\mathbf{A}_2', \quad k \geq 1, \tag{2.4}$$

and

$$\mathbf{\Sigma}_y \equiv \mathbf{\Sigma}_k(0) = \mathbf{A}_1\mathbf{\Sigma}_x\mathbf{A}_1' + \mathbf{A}_2\mathbf{\Sigma}_e\mathbf{A}_2'. \tag{2.5}$$

For a pre-specified integer $k_0 > 0$, define

$$\mathbf{M} = \sum_{k=1}^{k_0} \mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)', \tag{2.6}$$

which is a $p \times p$ semi-positive definite matrix. By $\mathbf{B}_1'\mathbf{A}_1 = \mathbf{0}$, we have $\mathbf{M}\mathbf{B}_1 = \mathbf{0}$, that is, the columns of $\mathbf{B}_1$ are the eigenvectors associated with the zero eigenvalues of $\mathbf{M}$, and the factor loading space $\mathcal{M}(\mathbf{A}_1)$ is spanned by the eigenvectors associated with the r non-zero eigenvalues of $\mathbf{M}$. For $k_0 > 1$, the summation in the definition of $\mathbf{M}$ enables us to pool information over different lags, which is particularly helpful when the sample size is small. In practice, with a given sample size, the estimation accuracy of auto-covariance matrices of $\mathbf{y}_t$ deteriorates as the lag k increases. Thus, some compromise in selecting k_0 is needed in a real application. Limited experience suggests that a relatively small k_0 is sufficient in providing useful information concerning the model structure of $\mathbf{y}_t$, because, for a stationary time series, cross-correlation matrices decay to zero exponentially as k increases. Also, the choice of k_0 seems to be not sensitive. See, for instance, the simulation results in Section 4. Note that the form of $\mathbf{M}$ in Equation (2.6) is a special case of the Orthonormalized Partial Least Squares of time series data. See the discussion in Section 5.

Turn to the estimation of the common factors. We observe that, from Equation (2.2),

$$\mathbf{B}_2'\mathbf{y}_t = \mathbf{B}_2'\mathbf{A}_1\mathbf{x}_t, \tag{2.7}$$

which is uncorrelated with $\mathbf{B}_1'\mathbf{y}_t$ defined in (2.3). Therefore,

$$\mathbf{B}_2'\mathbf{\Sigma}_y\mathbf{B}_1\mathbf{B}_1'\mathbf{\Sigma}_y\mathbf{B}_2 = \mathbf{0}, \tag{2.8}$$

which implies that $\mathbf{B}_2$ consists of the last r eigenvectors corresponding to the zero eigenvalues of $\mathbf{S} := \mathbf{\Sigma}_y\mathbf{B}_1\mathbf{B}_1'\mathbf{\Sigma}_y$. From the relationship in (2.7) and the discussion of Remark 1 in Section 3 below, $\mathbf{B}_2'\mathbf{A}_1$ is a $r \times r$ invertible matrix and hence $\mathbf{x}_t = (\mathbf{B}_2'\mathbf{A}_1)^{-1}\mathbf{B}_2'\mathbf{y}_t$. From Equation

(2.2), $\mathbf{x}_t$ does not include the white noise terms. Moreover, the columns in $\mathbf{A}_2$ can be treated as the eigenvectors associated with the non-zero eigenvalues of $\mathbf{S}$. Finally, even though $\mathbf{B}_2$ (also $\mathbf{A}_2$) is not unique and $\mathbf{B}_2\mathbf{H}$ is also a solution to (2.8) for any orthonormal matrix $\mathbf{H} \in R^{r\times r}$, this non-uniqueness does not alter the representation of $\mathbf{x}_t = (\mathbf{B}_2'\mathbf{A}_1)^{-1}\mathbf{B}_2'\mathbf{y}_t$.

Given the data $\{\mathbf{y}_t | t = 1, ..., n\}$, the first step in estimation is to estimate $\mathbf{A}_1$ or its column space $\mathcal{M}(\mathbf{A}_1)$, to recover the factor process $\mathbf{x}_t$, and to determine the number of common factors r. To begin, we assume for now that r is known. The estimation of r will be discussed later. Let $\widehat{\mathbf{\Sigma}}_y(k)$ be the lag-k sample auto-covariance matrix of $\mathbf{y}_t$. To estimate $\mathcal{M}(\mathbf{A}_1)$, we perform an eigen-analysis of

$$\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)'. \tag{2.9}$$

Let $\widehat{\mathbf{A}}_1 = (\widehat{\mathbf{a}}_1, ..., \widehat{\mathbf{a}}_r)$ and $\widehat{\mathbf{B}}_1 = (\widehat{\mathbf{b}}_1, ...\widehat{\mathbf{b}}_v)$ be two half orthonormal matrices consisting of the eigenvectors of $\widehat{\mathbf{M}}$ corresponding to the non-zero and zero eigenvalues, respectively. In view of Equation (2.8), we next perform another eigen-analysis of

$$\widehat{\mathbf{S}} = \widehat{\mathbf{\Sigma}}_y\widehat{\mathbf{B}}_1\widehat{\mathbf{B}}_1'\widehat{\mathbf{\Sigma}}_y, \tag{2.10}$$

which is a projected PCA. That is, we project the data $\mathbf{y}_t$ onto the direction of $\widehat{\mathbf{B}}_1$, then perform the PCA between the original data $\mathbf{y}_t$ and its projected coordinates. If the dimension p is small, we employ $\widehat{\mathbf{B}}_2 = (\widehat{\mathbf{b}}_{v+1}, ..., \widehat{\mathbf{b}}_p)$, where $\widehat{\mathbf{b}}_{v+1}, ..., \widehat{\mathbf{b}}_p$ are the eigenvectors corresponding to the smallest r eigenvalues of $\widehat{\mathbf{S}}$. On the other hand, if p is relatively large and the largest K eigenvalues of $\mathbf{\Sigma}_e$ are diverging, which is a reasonable condition in the high-dimensional case, we write $\mathbf{A}_2 = (\mathbf{A}_{21}, \mathbf{A}_{22})$ with $\mathbf{A}_{21} \in R^{p\times K}$ and $\mathbf{A}_{22} \in R^{p\times(v-K)}$ and consider the linear space $\mathcal{M}(\mathbf{B}_2^*)$, where $\mathbf{B}_2^* = (\mathbf{A}_{22}, \mathbf{B}_2) \in R^{p\times(p-K)}$. Note that $\mathbf{B}_2^*$ consists of $p-K$ eigenvectors corresponding to the $p-K$ smallest eigenvalues of $\mathbf{S} = \mathbf{\Sigma}_y\mathbf{B}_1\mathbf{B}_1'\mathbf{\Sigma}_y$ defined before. Let $\widehat{\mathbf{B}}_2^*$ be an estimator of $\mathbf{B}_2^*$ consisting of $p-K$ eigenvectors associated with the $p-K$ smallest eigenvalues of $\widehat{\mathbf{S}}$. We then estimate $\widehat{\mathbf{B}}_2$ by $\widehat{\mathbf{B}}_2 = \widehat{\mathbf{B}}_2^*\widehat{\mathbf{R}}$, where $\widehat{\mathbf{R}} = (\widehat{\mathbf{r}}_1, \ldots, \widehat{\mathbf{r}}_r) \in R^{(p-K)\times r}$ with $\widehat{\mathbf{r}}_i$ being the eigenvector associated with the i-th largest eigenvalues of $\widehat{\mathbf{B}}_2^{*\prime}\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1'\widehat{\mathbf{B}}_2^*$. This choice of estimator guarantees that the matrix $(\widehat{\mathbf{B}}_2'\widehat{\mathbf{A}}_1)^{-1}$ behaves well in recovering the common factor $\widehat{\mathbf{x}}_t$. Detailed properties of the estimators are given in Section 3. Finally, we recover the factor process as $\widehat{\mathbf{x}}_t = (\widehat{\mathbf{B}}_2'\widehat{\mathbf{A}}_1)^{-1}\widehat{\mathbf{B}}_2'\mathbf{y}_t$.

With $\widehat{\mathbf{A}}_1$ and the estimated factor process $\widehat{\mathbf{x}}_t$, we compute the h-step ahead prediction of

the $\mathbf{y}_t$ series using the formula $\widehat{\mathbf{y}}_{n+h} = \widehat{\mathbf{A}}_1 \widehat{\mathbf{x}}_{n+h}$, where $\widehat{\mathbf{x}}_{t+h}$ is an h-step ahead forecast for $\mathbf{x}_t$ based on the estimated past values $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n$. This can be done, for example, by fitting a VAR model to $\{\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n\}$.

2.3 Determination of the number of common factors

The estimation of $\mathbf{A}_1$ and $\mathbf{x}_t$ in the prior sections is based on a given r, which is unknown in practice. There are some methods available in the literature to determine r for the traditional factor model in Equation (1.1). See, for example, the information criterion in Bai and Ng (2002) and Bai (2003), the random matrix theory method in Onatski (2010), and the ratio-based method in Lam and Yao (2012), among others. However, none of these methods is applicable to Model (2.1) directly. The most relevant method is the one based on testing the number of zero canonical correlations between $\mathbf{y}_t$ and vectors of its lagged values employed in Gao and Tsay (2018). But this testing method only works when the dimension p is relatively small with respect to the sample size n.

In this section, we propose a new approach to estimate the number of common factors based on Equation (2.3), i.e. we perform white noise tests to determine the number of white noise components v and use $r = p - v$. Let $\widehat{\mathbf{G}}$ be the matrix of eigenvectors (in the decreasing order of eigenvalues) of the sample matrix $\widehat{\mathbf{M}}$ of Equation (2.9) and $\widehat{\mathbf{u}}_t = \widehat{\mathbf{G}}_t \mathbf{y}_t = (\widehat{u}_{1t}, \ldots, \widehat{u}_{pt})'$ be the transformed series. We propose to test sequentially the number of white noises in $\widehat{\mathbf{u}}_t$, which is an estimate of v. To this end, we consider two cases depending on the dimension p.

If the dimension p is small, we recommend using a bottom-up procedure to determine the number of white noise components. Specifically, we use the conventional test statistics, such as the well-known Ljung-Box statistic $Q(m)$ or its rank-based variant, to test the null hypothesis that $\widehat{u}_{it}$ is a white noise series starting with $i = p$. If the null hypothesis is rejected, then $\widehat{r} = i$; otherwise, reduce i by one and repeat the testing process. Clearly, this testing process can only last until $i = 1$. If all transformed series $\widehat{u}_{it}$ are white noise, then $\widehat{r} = 0$ and $\widehat{v} = p$. In general, if $\widehat{u}_{it}$ is not a white noise series but $\widehat{u}_{jt}$ are for $j = i+1, \ldots, p$, then $\widehat{r} = i$ and $\widehat{v} = p - i$, and we have $\widehat{\mathbf{G}} = [\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1]$, where $\widehat{\mathbf{A}}_1 \in R^{p \times \widehat{r}}$ and $\widehat{\mathbf{B}}_1 \in R^{p \times \widehat{v}}$.

For large p, the conventional white-noise test statistics are no longer adequate. But some methods have been developed in recent years to test high-dimensional white noise. We consider two such methods in this paper. The first method is introduced by Chang et al.

(2017) and makes use of the maximum absolute auto-correlations and cross-correlations of the component series. Specifically, let $\mathbf{w}_t = (w_{1t}, \ldots, w_{dt})'$ be a d-dimensional real-valued time series. In this paper, $1 \leq d \leq p$. Define the lag-k sample covariance matrix as $\widehat{\mathbf{\Sigma}}_w(k) = (n-k)^{-1} \sum_{t=k+1}^{n} (\mathbf{w}_t - \bar{\mathbf{w}})(\mathbf{w}_{t-k} - \bar{\mathbf{w}})'$, where $\bar{\mathbf{w}} = n^{-1} \sum_{t=1}^{n} \mathbf{w}_t$ is the sample mean. The test statistic T_n of Chang et al. (2017) is

$$T_n = \max_{1 \leq k \leq \bar{k}} T_{n,k}, \tag{2.11}$$

where $\bar{k} \geq 1$ is a pre-specified positive integer and $T_{n,k} = \max_{1 \leq j,l \leq d} n^{1/2} |\widehat{\rho}_{jl}(k)|$ with

$$\widehat{\mathbf{\Gamma}}_w(k) \equiv [\widehat{\rho}_{jl}(k)]_{1 \leq j,l \leq d} = \text{diag}\{\widehat{\mathbf{\Sigma}}_w(0)\}^{-1/2} \widehat{\mathbf{\Sigma}}_w(k) \text{diag}\{\widehat{\mathbf{\Sigma}}_w(0)\}^{-1/2}.$$

The limiting distribution of T_n can be approximated by that of the L_∞-norm of a normal random vector, i.e., there exists a random variable $\mathbf{Z}_d \sim N(\mathbf{0}, \mathbf{\Theta}_{d,n})$ such that

$$\sup_{s \geq 0} |P(T_n > s) - P(|\mathbf{Z}_d|_\infty > s)| = o(1),$$

where $\mathbf{\Theta}_{d,n}$ is the asymptotic covariance of the vector containing the columns of $\widehat{\mathbf{\Gamma}}_w(1)$ to $\widehat{\mathbf{\Gamma}}_w(\bar{k})$, and it can be estimated from $\{\mathbf{w}_t | t = 1, ..., n\}$. Therefore, the critical values of T_n can be obtained by bootstrapping from a multivariate normal distribution.

The second method of high-dimensional white noise test is introduced by Tsay (2018) using the extreme value theory. The test is simple with a close-form limiting distribution under some weak assumptions and is easy to use in practice. The basic idea of the test is as follows. Consider a d-dimensional time series $\mathbf{w}_t$ with a realization of n observations $\{\mathbf{w}_t | t = 1, \ldots, n\}$. Assume, for now, that $d < n$. Let $\widetilde{\mathbf{w}}_t = \mathbf{\Sigma}_w^{-1/2} \mathbf{w}_t$ be a standardized series, where $\mathbf{\Sigma}_w^{1/2}$ is a square-root matrix of the covariance matrix $\mathbf{\Sigma}_w$. With $d < n$, this standardization can be done by PCA. For simplicity, we denote the standardized realization as $\widetilde{\mathbf{w}}_t = \widehat{\mathbf{\Sigma}}_w^{-1/2} \mathbf{w}_t$. If $d \geq n$, $\widehat{\mathbf{\Sigma}}_w$ is singular and we discuss a modification later. Note that the components of $\widetilde{\mathbf{w}}_t = (\widetilde{w}_{1t}, \ldots, \widetilde{w}_{dt})'$ are mutually uncorrelated. Next, let $\widehat{\boldsymbol{\varrho}}_t = (\widehat{\varrho}_{1t}, \ldots, \widehat{\varrho}_{dt})'$ be the rank series of $\widetilde{\mathbf{w}}_t$, where $\widehat{\varrho}_{jt}$ is the rank of $\widetilde{w}_{jt}$ in $\{\widetilde{w}_{j,1}, ...\widetilde{w}_{j,n}\}$ for $1 \leq j \leq d$. The lag-ℓ rank cross-correlation matrix is then defined as

$$\widehat{\mathbf{\Gamma}}_{w,\ell} = \frac{12}{n(n^2-1)} \sum_{t=\ell+1}^{n} (\widehat{\boldsymbol{\varrho}}_t - \bar{\boldsymbol{\varrho}})(\widehat{\boldsymbol{\varrho}}_{t-\ell} - \bar{\boldsymbol{\varrho}})',$$

where $\bar{\boldsymbol{\varrho}} = \frac{n+1}{2}\mathbf{1}_d$ and $\mathbf{1}_d$ is a d-dimensional vector of ones. The test statistic of Tsay (2018) for testing that there is no serial or cross-sectional correlation in the first m lags of $\mathbf{w}_t$ is

$$T(m) = \max\{\sqrt{n}|\widehat{\boldsymbol{\Gamma}}_{w,\ell}(j,k)| : 1 \le j,k \le d,\ 1 \le \ell \le m\}, \tag{2.12}$$

where $\widehat{\boldsymbol{\Gamma}}_{w,\ell}(j,k)$ is the (j,k)-th element of $\widehat{\boldsymbol{\Gamma}}_{w,\ell}$. By the extreme-value theory, the limiting distribution of $T(m)$ under the white noise hypothesis is a Gumbel distribution provided that the component series of $\mathbf{w}_t$ follow a continuous distribution. Therefore, we reject the null hypothesis

$$H_0 : \mathbf{w}_t \text{ is a vector white noise,}$$

at the α-level if

$$T(m) \ge c_{d,m} \times x_{1-\alpha/2} + s_{d,m},$$

where $x_{1-\alpha/2} = -\log(-\log(1-\alpha/2))$ is the $(1-\alpha/2)$-th quantile of the standard Gumbel distribution and

$$c_{d,m} = [2\log(d^2m)]^{-1/2}, \text{ and } s_{d,m} = \sqrt{2\log(d^2m)} - \frac{\log(4\pi) + \log(\log(d^2m))}{2(2\log(d^2m))^{1/2}}.$$

When $d \ge n$, the sample covariance matrix of $\mathbf{w}_t$ is singular and some alternative methods must be sought to create mutually uncorrelated series. Tsay (2018) provided a method by selecting a subset series of $\mathbf{w}_t$ to perform testing, and the method works reasonably well in simulations and some applications. In this paper, we consider a simpler method by using the relations in (2.2) and (2.8). Note that in our testing, $\mathbf{w}_t$ is a subset of the transformed series $\widehat{\mathbf{u}}_t = \widehat{\mathbf{G}}'\mathbf{y}_t$. Since $\widehat{\mathbf{M}}$ is based on the covariance matrices of $\mathbf{y}_t$ and its lagged values the components of $\widehat{\mathbf{u}}_t$ associated with small eigenvalues contain little information on the dynamical dependence of $\mathbf{y}_t$. Therefore, we can drop the latter $(p - \varepsilon n)$ components from the transformed series without affecting the white-noise test, where $\varepsilon \in (0,1)$. In other words, when $p > n$, we cannot start with $\mathbf{w}_t = \widehat{\mathbf{u}}_t$, but we can choose $\mathbf{w}_t$ to consist of the first $d = \varepsilon n < n$ components of $\widehat{\mathbf{u}}_t$ to perform the white-noise test without affecting the determination of r under the assumption that r is small in applications.

Return to the determination of r when p is large. We can apply the high-dimensional white noise test of Chang et al. (2017) or Tsay (2018) to subsets of the transformed series $\widehat{\mathbf{u}}_t$. Specifically, let $p_* = p$ if $p < n$ and $p_* = \varepsilon n$ if $p \ge n$, where $\varepsilon \in (0,1)$. Starting with $i = 1$ and $\mathbf{w}_t = (\widehat{u}_{i,t}, \ldots, \widehat{u}_{p_*,t})'$, we test the null hypothesis that $\mathbf{w}_t$ has no serial or cross-sectional

correlations in the first m lags using a selected test statistic. If the null hypothesis is rejected, increase i by one and repeat the testing process. Using this testing process, we select $\widehat{r}$ as $i-1$ for which the ith test does not reject the null hypothesis. Note that since both test statistics considered use the maximum of absolute correlations, the computation of the testing process is trivial because we only need to compute the cross-correlation matrices of $\widehat{\mathbf{u}}_t$ once.

3 Theoretical Properties

In this section, we investigate the asymptotic theory for the estimation method used in the paper. Starting with the assumption that the number of common factors r is known, we divide the derivations into two cases depending on the value of the dimension p. The case of estimated r is discussed later.

3.1 Asymptotic properties when $n \to \infty$ and p is fixed

We consider first the asymptotic properties of the estimators when p is fixed but $n \to \infty$. These properties show the behavior of our estimation method when n is large and p is relatively small. We begin with the assumptions used.

Assumption 1. *The process $\{(\mathbf{y}_t, \mathbf{f}_t)\}$ is α-mixing with the mixing coefficient satisfying the condition $\sum_{k=1}^{\infty} \alpha_p(k)^{1-2/\gamma} < \infty$ for some $\gamma > 2$, where*

$$\alpha_p(k) = \sup_i \sup_{A \in \mathcal{F}_{-\infty}^i, B \in \mathcal{F}_{i+k}^{\infty}} |P(A \cap B) - P(A)P(B)|,$$

and $\mathcal{F}_i^j$ is the σ-field generated by $\{(\mathbf{y}_t, \mathbf{f}_t) : i \leq t \leq j\}$.

Assumption 2. *$E|f_{it}|^{2\gamma} < C_1$ and $E|\varepsilon_{jt}|^{2\gamma} < C_2$ for $1 \leq i \leq r$ and $1 \leq j \leq v$, where $C_1, C_2 > 0$ are some constants and γ is given in Assumption 1.*

Assumption 3. *$\lambda_1 > ... > \lambda_r > \lambda_{r+1} = ... = \lambda_p = 0$, where λ_i is the i-th largest eigenvalue of $\mathbf{M}$ in Equation (2.6).*

Assumption 1 is standard for dependent random processes. See Gao et al. (2018) for a theoretical justification for VAR models. The conditions in Assumption 2 imply that $E|y_{it}|^{2\gamma} < C$ under the setting that p is fixed. In Assumption 3, if the r non-zero eigenvalues of $\mathbf{M}$ are distinct, the eigenvector matrix $\mathbf{A}_1$ is uniquely defined if we ignore the trivial

replacement of $\mathbf{a}_j$ by $-\mathbf{a}_j$ for $1 \leq j \leq r$. The following theorem establishes the consistency of the estimated loading matrix $\widehat{\mathbf{A}}_1$, its orthonormal complement $\widehat{\mathbf{B}}_1$, the matrix $\widehat{\mathbf{B}}_2$ and the extracted common factor $\widehat{\mathbf{A}}_1\widehat{\mathbf{x}}_t$.

Theorem 1. *Suppose Assumptions 1-3 hold and r is known and fixed. Then, for fixed p,*

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(n^{-1/2}), \|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2 = O_p(n^{-1/2}) \text{ and } \|\widehat{\mathbf{B}}_2 - \mathbf{B}_2\|_2 = O_p(n^{-1/2}),$$

as $n \to \infty$. Therefore,

$$\|\widehat{\mathbf{A}}_1\widehat{\mathbf{x}}_t - \mathbf{A}_1\mathbf{x}_t\|_2 = O_p(n^{-1/2}).$$

Remark 1. *From Theorem 1 and as expected, the convergence rates of all estimates are standard at $\sqrt{n}$, which is commonly seen in the traditional statistical theory. To recover the factor process, we need to guarantee that $\mathbf{B}_2'\mathbf{A}_1$ is invertible. This follows from the fact that there exist $\mathbf{R}_1 \in R^{r\times r}$ and $\mathbf{R}_2 \in R^{v\times r}$ such that $\mathbf{B}_2 = \mathbf{L}_1\mathbf{R}_1 + \mathbf{L}_2\mathbf{R}_2 = \mathbf{A}_1\mathbf{Q}_1\mathbf{R}_1 + \mathbf{A}_2\mathbf{Q}_2\mathbf{R}_2$, i.e. each column of $\mathbf{B}_2$ can be represented as a linear combination of the columns of $\mathbf{L}$. Therefore, $\mathbf{I}_r = \mathbf{B}_2'\mathbf{B}_2 = \mathbf{B}_2'\mathbf{A}_1\mathbf{Q}_1\mathbf{R}_1$ and hence rank$(\mathbf{B}_2'\mathbf{A}_1) = r$ which is of full rank.*

In general, the choice of $\mathbf{A}_1$ in Model (2.2) is not unique so we consider the error in estimating $\mathcal{M}(\mathbf{A}_1)$, the column space of $\mathbf{A}_1$, because $\mathcal{M}(\mathbf{A}_1)$ is uniquely defined by (2.2) and it does not vary with different choices of $\mathbf{A}_1$. The same argument also applies to matrices $\mathbf{A}_2$, $\mathbf{B}_1$ and $\mathbf{B}_2$. To this end, we adopt the discrepancy measure used by Pan and Yao (2008): for two $p \times r$ half orthogonal matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ satisfying the condition $\mathbf{H}_1'\mathbf{H}_1 = \mathbf{H}_2'\mathbf{H}_2 = \mathbf{I}_r$, the difference between the two linear spaces $\mathcal{M}(\mathbf{H}_1)$ and $\mathcal{M}(\mathbf{H}_2)$ is measured by

$$D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) = \sqrt{1 - \frac{1}{r}\text{tr}(\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_2\mathbf{H}_2')}. \tag{3.1}$$

Note that $D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) \in [0, 1]$. It is equal to 0 if and only if $\mathcal{M}(\mathbf{H}_1) = \mathcal{M}(\mathbf{H}_2)$, and to 1 if and only if $\mathcal{M}(\mathbf{H}_1) \perp \mathcal{M}(\mathbf{H}_2)$. The following theorem establishes the convergence of $D(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{A}_1))$ when $\mathbf{A}_1$ is not uniquely defined.

Theorem 2. *Suppose Assumptions 1-2 hold and r is known and fixed. Then, for fixed p,*

$$D(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{A}_1)) = O_p(n^{-1/2}), \quad D(\mathcal{M}(\widehat{\mathbf{B}}_1), \mathcal{M}(\mathbf{B}_1)) = O_p(n^{-1/2})$$

and

$$D(\mathcal{M}(\widehat{\mathbf{B}}_2), \mathcal{M}(\mathbf{B}_2)) = O_p(n^{-1/2}),$$

as $n \to \infty$. The convergence rate of the extracted factors $\widehat{\mathbf{A}}_1\widehat{\mathbf{x}}_t$ is the same as that in Theorem 1.

This theorem implies that the convergence rate does not change even when some non-zero eigenvalues for $\mathbf{M}$ are not distinct and $\mathbf{A}_1$ is not uniquely defined. In fact, the consistency of the linear spaces of $\mathcal{M}(\mathbf{B}_1)$ and $\mathcal{M}(\mathbf{B}_2)$ is more meaningful since their columns correspond to the zero eigenvalues of $\mathbf{M}$ and $\mathbf{S}$, respectively, and they cannot be uniquely characterized.

3.2 Asymptotic properties when $n \to \infty$ and $p \to \infty$

Turn to the case of high-dimensional time series. It is well known that if the dimension p diverges faster than $n^{1/2}$, the sample covariance matrix is no longer a consistent estimate of the population covariance matrix. When $p = o(n^{1/2})$, it is still possible to consistently estimate the factor loading matrix $\mathbf{A}$ and the number of common factors r. See Gao and Tsay (2018) for details. Therefore, without any additional assumptions on the underlying structure of time series, p can only be as large as $o(n^{1/2})$. To deal with the case of large p, we impose some conditions on the transformation matrix $\mathbf{L}$ of Equation (2.1) and the cross dependence of $\mathbf{y}_t$.

Let $\mathbf{L} = (\mathbf{c}_1, ..., \mathbf{c}_p)$, where $\mathbf{c}_i$ is a p-dimensional column vector, and hence, $\mathbf{L}_1 = (\mathbf{c}_1, ..., \mathbf{c}_r)$ and $\mathbf{L}_2 = (\mathbf{c}_{r+1}, ..., \mathbf{c}_p)$.

Assumption 4. *(i) $\mathbf{L}_1 = (\mathbf{c}_1, ..., \mathbf{c}_r)$ such that $\|\mathbf{c}_j\|_2^2 \asymp p^{1-\delta_1}$, $j = 1, ..., r$ and $\delta_1 \in [0, 1)$; (ii) For each $j = 1, ..., r$ and δ_1 given in (i), $\min_{\theta_i \in R, i \neq j} \|\mathbf{c}_j - \sum_{1 \leq i \leq r, i \neq j} \theta_i \mathbf{c}_i\|_2^2 \asymp p^{1-\delta_1}$.*

Assumption 5. *(i) $\mathbf{L}_2$ admits a singular value decomposition $\mathbf{L}_2 = \mathbf{A}_2\mathbf{D}_2\mathbf{V}_2'$, where $\mathbf{A}_2 \in R^{p \times v}$ is given in Equation (2.2), $\mathbf{D}_2 = diag(d_1, ..., d_v)$ and $\mathbf{V}_2 \in R^{v \times v}$ satisfying $\mathbf{V}_2'\mathbf{V}_2 = \mathbf{I}_v$; (ii) There exists a finite integer $0 < K < v$ such that $d_1 \asymp ... \asymp d_K \asymp p^{(1-\delta_2)/2}$ for some $\delta_2 \in [0, 1)$ and $d_{K+1} \asymp ... \asymp d_v \asymp 1$.*

Assumption 6. *$0 \leq \kappa_{\min} \leq \|\mathbf{\Sigma}_{f\varepsilon}(k)\|_2 \leq \kappa_{\max}$ for $1 \leq k \leq k_0$, where $\kappa_{\min}$ and $\kappa_{\max}$ can be either finite constants or diverging rates in relation to p and n.*

Assumption 7. *(i) For any $\mathbf{h} \in R^v$ with $\|\mathbf{h}\|_2 = 1$, $E|\mathbf{h}'\boldsymbol{\varepsilon}_t|^{2\gamma} < \infty$; (ii) $\sigma_{\min}(\mathbf{R}'\mathbf{B}_2^{*\prime}\mathbf{A}_1) \geq C_3$ for some constant $C_3 > 0$ and some half orthogonal matrix $\mathbf{R} \in R^{(p-K) \times r}$ satisfying $\mathbf{R}'\mathbf{R} = \mathbf{I}_r$, where $\sigma_{\min}$ denotes the minimum non-zero singular value of a matrix.*

The quantity δ_1 of Assumption 4 is used to quantify the strength of the factors. If $\delta_1 = 0$, the corresponding factors are called strong factors, since it includes the case where

each element of $\mathbf{c}_i$ is $O(1)$. If $\delta_1 > 0$, the corresponding factors are weak factors and the smaller the δ_1 is, the stronger the factors are. One advantage of using index δ_1 is to link the convergence rates of the estimated factors explicitly to the strength of the factors. Assumption 4 ensures that all common factors in $\mathbf{x}_t$ are of equal strength δ_1. There are many sufficient conditions for Assumption 5 to hold. For example, it holds if we allow $(\mathbf{c}_{r+1}, ..., \mathbf{c}_{r+K})$ to satisfy Assumption 4, and the L_1- and L_∞-norms of $(\mathbf{c}_{r+K+1}, ..., \mathbf{c}_p)$ are all finite. A special case is to let $\mathbf{c}_{r+K+j}$ be a standard unit vector. In Assumption 6, $\kappa_{\min}$ and $\kappa_{\max}$ control the strength of the dependence between $\mathbf{f}_t$ and the past errors $\boldsymbol{\varepsilon}_{t-j}$ for $j \geq 1$. The maximal order of $\kappa_{\max}$ is $p^{1/2}$ which is the Frobenius norm of $\boldsymbol{\Sigma}_{f\varepsilon}(k)$ and $\kappa_{\max} = 0$ (hence $\kappa_{\min} = 0$) implies that $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_s$ are independent for all t and s. Throughout this article, if $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_s$ are independent for all t and s, then $\kappa_{\min} = \kappa_{\max} = 0$ and all the conditions and expressions below involved with $\kappa_{\min}$ and $\kappa_{\max}$ will be removed. Assumption 7(i) is mild and includes the standard normal distribution as a special case. Assumption 7(ii) is reasonable since $\mathbf{B}_2$ is a subspace of $\mathbf{B}_2^*$, and Remark 1 implies that $\mathbf{R}'\mathbf{B}_2^{*\prime}\mathbf{A}_1$ is invertible. The choice of $\widehat{\mathbf{R}}$ and hence $\widehat{\mathbf{B}}_2 = \widehat{\mathbf{B}}_2^*\widehat{\mathbf{R}}$ will be discussed later.

Remark 2. *In Assumption 5, we actually only require $d_K \asymp p^{(1-\delta_2)/2}$ for some $\delta_2 \in [0, 1)$ and $K \geq 1$, and the upper singular values $\{d_1, ..., d_{K-1}\}$ if $K > 1$ can be even larger provided that the largest one d_1 should be bounded by another rate $p^{(1-\delta_3)/2}$ for some $0 \leq \delta_3 \leq \delta_2$. For simplicity, we assume the top singular values are of the same order.*

If p is large, it is not possible to consistently estimate $\mathbf{B}_2$ or even $\mathcal{M}(\mathbf{B}_2)$. Instead, we will estimate $\mathbf{B}_2^* = (\mathbf{A}_{22}, \mathbf{B}_2)$ or equivalently $\mathcal{M}(\mathbf{B}_2^*)$, which is the subspace spanned by the eigenvectors associated with the $p - K$ smallest eigenvalues of $\mathbf{S}$. Assume $\widehat{\mathbf{B}}_2^*$ consists of the eigenvectors corresponding to the smallest $p-K$ eigenvalues of $\widehat{\mathbf{S}}$. Under some conditions, we can show that $\mathcal{M}(\widehat{\mathbf{B}}_2^*)$ is consistent to $\mathcal{M}(\mathbf{B}_2^*)$. This is also the case in the literature on high-dimensional PCA with i.i.d. data. See, for example, Shen et al. (2016) and the references therein. Therefore, the choice of $\widehat{\mathbf{B}}_2$ should be a subspace of $\widehat{\mathbf{B}}_2^*$, and we will discuss it before Theorem 5 below.

Theorem 3. *Suppose Assumptions 1-7 hold and r is known and fixed. As $n \to \infty$, if $p^{\delta_1} n^{-1/2} = o(1)$ or $\kappa_{\max}^{-1} p^{\delta_1/2+\delta_2/2} n^{-1/2} = o(1)$, then*

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = \begin{cases} O_p(p^{\delta_1} n^{-1/2}), & \text{if } \kappa_{\max} p^{\delta_1/2-\delta_2/2} = o(1), \\ O_p(\kappa_{\min}^{-2} p^{\delta_2} n^{-1/2} + \kappa_{\min}^{-2}\kappa_{\max} p^{\delta_1/2+\delta_2/2} n^{-1/2}), & \text{if } r \leq K, \kappa_{\min}^{-1} p^{\delta_2/2-\delta_1/2} = o(1), \\ O_p(\kappa_{\min}^{-2} p n^{-1/2} + \kappa_{\min}^{-2}\kappa_{\max} p^{1+\delta_1/2-\delta_2/2} n^{-1/2}), & \text{if } r > K, \kappa_{\min}^{-1} p^{(1-\delta_1)/2} = o(1), \end{cases}$$

and the above results also hold for $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2$. *Furthermore,*

$$\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p\left(p^{2\delta_2-\delta_1}n^{-1/2} + p^{\delta_2}n^{-1/2} + (1+p^{2\delta_2-2\delta_1})\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2\right).$$

Remark 3. *(i) If* $\kappa_{max} = \kappa_{\min} = 0$, *i.e.,* $\mathbf{f}_t$ *and* $\boldsymbol{\varepsilon}_s$ *are independent for all* t *and* s, *we have*

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(p^{\delta_1}n^{-1/2}) \text{ and } \|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p(p^{2\delta_2-\delta_1}n^{-1/2} + p^{\delta_2}n^{-1/2} + p^{\delta_1}n^{-1/2}).$$

To guarantee these estimates are consistent, we require $p^{\delta_1}n^{-1/2} = o(1)$, $p^{\delta_2}n^{-1/2} = o(1)$ *and* $p^{2\delta_2-\delta_1}n^{-1/2} = o(1)$. *When* $p \asymp n^{1/2}$, *it implies that* $0 \le \delta_1 < 1$, $0 \le \delta_2 < 1$ *and* $\delta_2 < (1+\delta_1)/2$, *i.e., the ranges of* δ_1 *and* δ_2 *are pretty wide. On the other hand, if* $p \asymp n$, *we see that* $0 \le \delta_1 < 1/2$, $0 \le \delta_2 < 1/2$ *and* $2\delta_2 - \delta_1 < 1/2$, *these ranges become narrower if* p *is large.*
(ii) When $\kappa_{\max} \neq 0$ *and* $\kappa_{\min} \neq 0$, *there are many possible results. A reasonable assumption is* $\kappa_{\min} \asymp \kappa_{\max} \asymp p^{\delta/2}$ *for some* $0 \le \delta < 1$ *since* r *is small. For example, set* $\delta = \delta_1$,

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = \begin{cases} O_p(p^{\delta_1}n^{-1/2}), & \text{if } p^{\delta_1-\delta_2/2} = o(1), \\ O_p(p^{\delta_2/2}n^{-1/2}), & \text{if } r \le K, \kappa_{\min}^{-1}p^{\delta_2/2-\delta_1/2} = o(1), \end{cases}$$

and there is no consistency result when $r > K$. *Furthermore we have* $\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p(p^{2\delta_2-\delta_1}n^{-1/2} + p^{\delta_2}n^{-1/2})$. *Thus, we require* $p^{\delta_1}n^{-1/2} = o(1)$, $p^{\delta_2}n^{-1/2} = o(1)$ *and* $p^{2\delta_2-\delta_1}n^{-1/2} = o(1)$. *The ranges of* δ_1 *and* δ_2 *are the same as discussed in Remark 3(i) above, we omit the details here. On the other hand, if* $\delta > (1-\delta_1)/2$, *it is still possible to consistently estimate them when* $r > K$, *the discussion is similar and is omitted for simplicity.*

From Theorem 3, we see that when $p \asymp n$, we require $\delta_1 < 1/2$ and $\delta_2 < 1/2$ to guarantee the consistency of our estimation method, which rules out the cases of the presence of weaker factors with $\delta_1 \ge 1/2$ and a slower diverging of the noise covariance matrix with $\delta_2 \ge 1/2$. The convergence rates in Theorem 3 are not optimal and they can be further improved under additional assumption on $\boldsymbol{\varepsilon}_t$ below.

Assumption 8. *For any* $\mathbf{h} \in R^v$ *with* $\|\mathbf{h}\|_2 = 1$, *there exists a constant* $C_4 > 0$ *such that*

$$P(|\mathbf{h}'\boldsymbol{\varepsilon}_t| > x) \le 2\exp(-C_4x^2) \quad \text{for any } x > 0.$$

Assumption 8 implies that $\boldsymbol{\varepsilon}_t$ are sub-Gaussian. Examples of sub-Gaussian distributions include the standard normal distribution in R^v, the uniform distribution on the cube $[-1,1]^v$,

among others. See, for example, Vershynin (2018).

Theorem 4. *Let Assumptions 1-8 hold and r is known and fixed, and $p^{\delta_1/2}n^{-1/2} = o(1)$, $p^{\delta_2/2}n^{-1/2} = o(1)$.*

(i) Under the condition that $\delta_1 \leq \delta_2$,

$$\|\widehat{\mathbf{A}}_1-\mathbf{A}_1\|_2 = \begin{cases} O_p(p^{\delta_1/2}n^{-1/2}), & \text{if } \kappa_{\max}p^{\delta_1/2-\delta_2/2} = o(1), \\ O_p(\kappa_{\min}^{-2}p^{\delta_2-\delta_1/2}n^{-1/2} + \kappa_{\min}^{-2}\kappa_{\max}p^{\delta_2/2}n^{-1/2}), & \text{if } r \leq K, \kappa_{\min}^{-1}p^{\delta_2/2-\delta_1/2} = o(1), \\ O_p(\kappa_{\min}^{-2}p^{1-\delta_1/2}n^{-1/2} + \kappa_{\min}^{-2}\kappa_{\max}p^{1-\delta_2/2}n^{-1/2}), & \text{if } r > K, \kappa_{\min}^{-1}p^{(1-\delta_1)/2} = o(1), \end{cases}$$

and the above results also hold for $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2$, and

$$\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p(p^{2\delta_2-3\delta_1/2}n^{-1/2} + p^{2\delta_2-2\delta_1}\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2).$$

(ii) Under the condition that $\delta_1 > \delta_2$, if $\kappa_{\max} = 0$ and $p^{\delta_1-\delta_2/2}n^{-1/2} = o(1)$, then

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(p^{\delta_1-\delta_2/2}n^{-1/2}).$$

If $\kappa_{\max} >> 0$, then

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = \begin{cases} O_p(\kappa_{\min}^{-2}\kappa_{\max}p^{\delta_1/2}n^{-1/2}), & \text{if } r \leq K, \kappa_{\min}^{-1}p^{\delta_2/2-\delta_1/2} = o(1), \\ O_p(\kappa_{\min}^{-2}\kappa_{\max}p^{1+\delta_1/2-\delta_2}n^{-1/2}), & \text{if } r > K, \kappa_{\min}^{-1}p^{(1-\delta_1)/2} = o(1), \end{cases}$$

and the above results also hold for $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2$, and

$$\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2 = O_p(p^{\delta_2/2}n^{-1/2} + \|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2).$$

Remark 4. *(i) Consider the case $\kappa_{\min} = \kappa_{\max} = 0$. If $\delta_1 \leq \delta_2$, $\|\widehat{\mathbf{A}}_1-\mathbf{A}_1\|_2 = O_p(p^{\delta_1/2}n^{-1/2})$ and $\|\widehat{\mathbf{B}}_2^*-\mathbf{B}_2^*\|_2 = O_p(p^{2\delta_2-3\delta_1/2}n^{-1/2})$. For $p \asymp n$, we require $0 \leq \delta_1 \leq \delta_2 < 1$ and $4\delta_2-3\delta_1 < 1$, or equivalently $0 \leq \delta_1 \leq \delta_2 < 3\delta_1/4 + 1/4$. If $\delta_1 > \delta_2$, $\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 = O_p(p^{\delta_1-\delta_2/2}n^{-1/2})$ and $\|\widehat{\mathbf{B}}_2^*-\mathbf{B}_2^*\|_2 = O_p(p^{\delta_2/2}n^{-1/2})$. Thus, if $p \asymp n$, we require $\max\{2\delta_1-1, 0\} < \delta_2 < \delta_1 < 1$. Therefore, if $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_s$ are independent and $p \asymp n$, δ_1 and δ_2 need to satisfy $0 \leq \delta_1 \leq \delta_2 < 3\delta_1/4+1/4$ or $\max\{2\delta_1-1, 0\} < \delta_2 < \delta_1 < 1$, which is much wider than those of Theorem 3. (ii) If $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_s$ are correlated for $s < t$, we may have many consistency results depending on the strength of the dependence between $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_s$. We omit the details here.*

When $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{B}_1$ and $\mathbf{B}_2$ are not uniquely defined, we can still have similar results as

Theorem 2 by replacing the rates with their counterparts in Theorem 3 or Theorem 4 under different conditions. For simplicity, we do not repeat the results here. Once we have $\widehat{\mathbf{B}}_2^*$, we suggest to choose $\widehat{\mathbf{B}}_2$ as $\widehat{\mathbf{B}}_2 = \widehat{\mathbf{B}}_2^*\widehat{\mathbf{R}}$, where $\widehat{\mathbf{R}} = (\widehat{\mathbf{r}}_1, .., \widehat{\mathbf{r}}_r) \in R^{(p-K)\times r}$, and $\widehat{\mathbf{r}}_i$ is the vector associated with the i-th largest eigenvalues of $\widehat{\mathbf{B}}_2^{*\prime}\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1'\widehat{\mathbf{B}}_2^*$. This choice can guarantee that the matrix $(\widehat{\mathbf{B}}_2'\widehat{\mathbf{A}}_1)^{-1}$ behaves well when recovering the factor $\widehat{\mathbf{x}}_t$. On the other hand, this choice could still eliminate the diverging part of the noise covariance matrix and gives prominent convergence rate, as shown in Theorem 5. There are many ways to choose the number of components K in Assumption 5 so long as $p - K > r$. We will discuss the choice of K in Remark 5 below and also in Section 5. The following theorem states the convergence rate of the extracted common factors.

Theorem 5. *Under the conditions in Theorem 3 or 4, we have*

$$p^{-1/2}\|\widehat{\mathbf{A}}_1\widehat{\mathbf{x}}_t - \mathbf{A}_1\mathbf{x}_t\|_2 = O_p(p^{-1/2} + p^{-\delta_1/2}\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_2 + p^{-\delta_2/2}\|\widehat{\mathbf{B}}_2^* - \mathbf{B}_2^*\|_2).$$

Remark 5. *(i) A similar result is given in Theorem 3 of Lam et al. (2011), which deals with the approximate factor models. When $\delta_1 = \delta_2 = 0$, i.e. the factors and the noise terms are all strong, the convergence rate in Theorem 5 is $O_p(p^{-1/2} + n^{-1/2})$, which is the optimal rate specified in Theorem 3 of Bai (2003) when dealing with the traditional approximate factor models.*

(ii) It is a common issue to select the number of principle components in the literature and there are many possible approaches available. Since it is impossible to eliminate all the noise effects in recovering the factors and we just need to guarantee that the diverging part of the noises are removed for large p. Thus, we may select K in a range of possible values. In practice, Let $\widehat{\mu}_1 \geq ... \geq \widehat{\mu}_p$ be the sample eigenvalues of $\widehat{\mathbf{S}}$ and define $\widehat{K}_L$ as

$$\widehat{K}_L = \arg \min_{1\leq j\leq \widehat{K}_U} \{\widehat{\mu}_{j+1}/\widehat{\mu}_j\}, \tag{3.2}$$

and $\widehat{K}_U$ is a pre-specified integer. In practice, we suggest $\widehat{K}_U = \min\{\sqrt{p}, \sqrt{n}, p-\widehat{r}, 10\}$. Then the estimator $\widehat{K}$ for K can assume some value between $\widehat{K}_L$ and $\widehat{K}_U$.

Next, we study the consistency of the white noise test described in Section 2. In fact, the consistency conditions depend on which method we use. We only present the consistency when p is large since the case of small p is trivial.

Theorem 6. *(i) Let Assumptions 1-8 hold. If $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2^2\|\mathbf{\Sigma}_y\|_2 = o_p(1)$, then the test*

statistic T_n defined in (2.11) can consistently estimate r, i.e. $P(\widehat{r} = r) \to 1$ as $n \to \infty$.
(ii) Let Assumptions 1-8 hold. If $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_2 \|\mathbf{\Sigma}_y\|_2 = o_p(1)$ and $\|\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_y\|_2 = o_p(1)$, the test statistic $T(m)$ defined in (2.12) can also consistently estimate r.

Remark 6. *(i) If $\mathbf{f}_t$ and $\boldsymbol{\varepsilon}_s$ are independent for all t and s, the conditions in Theorem 6(i) are essentially $pn^{-1} = o(1)$ if $\delta_1 \leq \delta_2$ and $p^{1+2\delta_1-2\delta_2}n^{-1} = o(1)$ if $\delta_1 > \delta_2$. Thus, we require $p \asymp n^{\xi}$ for $0 < \xi < 1$ for both cases. As for Theorem 6(ii), the conditions are $p^{1-\delta_1/2}n^{-1/2} = o(1)$ if $\delta_1 \leq \delta_2$ and $p^{1+\delta_1-3\delta_2/2}n^{-1/2} = o(1)$ if $\delta_1 > \delta_2$. We also require $p \asymp n^{\xi}$ for some $0 < \xi < 1$. However the conditions in Theorem 6(ii) are slightly stronger than those in Theorem 6(i) since we need to establish the consistency of the covariance matrix while (i) does not.*

(ii) Even though the conditions in Theorem 6(ii) is slightly stronger, the method based on $T(m)$ is simple and easy to use, and the performance is also satisfactory when p is moderately large. See Tsay (2018) and the simulation results in Section 4 for details.

With the estimator $\widehat{r}$, we may define the estimator for $\mathbf{A}_1$ as $\widehat{\mathbf{A}}_1 = (\widehat{\mathbf{a}}_1, ..., \widehat{\mathbf{a}}_{\widehat{r}})$, where $\widehat{\mathbf{a}}_1, ..., \widehat{\mathbf{a}}_{\widehat{r}}$ are the orthonormal eigenvectors of $\widehat{\mathbf{M}}$, defined in (2.9), corresponding to the $\widehat{r}$ largest eigenvalues. In addition, we may also replace r by $\widehat{r}$ in the whole methodology described in Section 2.

4 Numerical Properties

4.1 Simulation

In this section, we illustrate the finite-sample properties of the proposed methodology under the scenarios when p is small and large, respectively. As the dimensions of $\widehat{\mathbf{A}}_1$ and $\mathbf{A}_1$ are not necessarily the same, and $\mathbf{L}_1$ is not an orthogonal matrix in general, we first extend the discrepancy measure in Equation (3.1) to a more general form below. Let $\mathbf{H}_i$ be a $p \times r_i$ matrix with $\text{rank}(\mathbf{H}_i) = r_i$, and $\mathbf{P}_i = \mathbf{H}_i(\mathbf{H}_i'\mathbf{H}_i)^{-1}\mathbf{H}_i'$, $i = 1, 2$. Define

$$\bar{D}(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) = \sqrt{1 - \frac{1}{\min(r_1, r_2)}\text{tr}(\mathbf{P}_1\mathbf{P}_2)}. \tag{4.1}$$

Then $\bar{D} \in [0, 1]$. Furthermore, $\bar{D}(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) = 0$ if and only if either $\mathcal{M}(\mathbf{H}_1) \subset \mathcal{M}(\mathbf{H}_2)$ or $\mathcal{M}(\mathbf{H}_2) \subset \mathcal{M}(\mathbf{H}_1)$, and it is 1 if and only if $\mathcal{M}(\mathbf{H}_1) \perp \mathcal{M}(\mathbf{H}_2)$. When $r_1 = r_2 = r$ and

$\mathbf{H}_i'\mathbf{H}_i = \mathbf{I}_r$, $\bar{D}(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2))$ is the same as that in Equation (3.1). We only present the simulation results by taking $k_0 = 2$ in Equation (2.9) to save space since other choices of k_0 produce similar patterns.

Example 1. Consider Model (2.1) with common factors satisfying

$$\mathbf{f}_t = \mathbf{\Phi}\mathbf{f}_{t-1} + \boldsymbol{\eta}_t,$$

where $\boldsymbol{\eta}_t$ is a white noise process. We set the true number of factors $r = 3$, the dimension $p = 5, 10, 15, 20$, and the sample size $n = 200, 500, 1000, 1500, 3000$, respectively. For each realization, the elements of $\mathbf{L}$ are drawn independently from $U(-2, 2)$, and the elements of $\mathbf{L}_2$ are then divided by $\sqrt{p}$ to balance the accumulated variances of f_{it} and ε_{it} for each component of $\mathbf{y}_t$. $\mathbf{\Phi}$ is a diagonal matrix with its diagonal elements being drawn independently from $U(0.5, 0.9)$, $\boldsymbol{\varepsilon}_t \sim N(0, \mathbf{I}_v)$ and $\boldsymbol{\eta}_t \sim N(0, \mathbf{I}_r)$. We use 1000 replications for each (p, n) configuration.

We first study the performance of the estimation of the number of factors. Since p is relatively small compared to the sample size n, for each iteration, we use Ljung-Box test statistics with $m = 10$ to determine the number of factors, i.e. $Q(10)$. The empirical probabilities $P(\widehat{r} = r)$ are reported in Table 1. From the table, we see that, for each given p, the performance of the proposed method improves as the sample size increases. On the other hand, for a given n, the proportion of the empirical probability decreases slightly as p increases, which is reasonable since it is harder to determine the correct number of factors when the dimension increases and the errors in the testing procedure accumulates. Overall, the Ljung-Box test works well for the case of small dimension (e.g., $p \leq 10$). However, when p is slightly larger (e.g., $p = 15, 20$), the test statistic tends to overestimate the number of factors, implying that we can still keep sufficient information of the original process $\mathbf{y}_t$. To illustrate this, we present the boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{L}_1))$ in Figure 1, where $\bar{D}(\cdot, \cdot)$ is defined in (4.1). From Figure 1, for each p, the discrepancy decreases as the sample size increases and this is in agreement with our theory.

Furthermore, for each (p, n), we study the root-mean-square error (RMSE):

$$\text{RMSE} = \left(\frac{1}{n} \sum_{t=1}^{n} \|\widehat{\mathbf{A}}_1 \widehat{\mathbf{x}}_t - \mathbf{L}_1 \mathbf{f}_t\|_2^2 \right)^{1/2}, \tag{4.2}$$

Table 1: Empirical probabilities $P(\widehat{r} = r)$ of various (p, n) configurations for the model of Example 1 with $r = 3$, where p and n are the dimension and the sample size, respectively. 1000 iterations are used.

		n				
	p	200	500	1000	1500	3000
$r = 3$	5	0.861	0.889	0.890	0.912	0.926
	10	0.683	0.718	0.723	0.735	0.748
	15	0.506	0.555	0.561	0.599	0.601
	20	0.395	0.425	0.441	0.447	0.453

which quantifies the accuracy in estimating the common factor process. Boxplots of the RMSE are shown in Figure 2. From the plot, we see a clear pattern that, as the sample size increases, the RMSE decreases for each p, which is consistent with the results of Theorem 1. Overall, the one-by-one testing procedure works well when the sample size is small, and the RMSE is decreasing when the sample size increases, even though the performance of the test may deteriorate due to the overestimation of the number of the factors for higher dimension p.

Example 2. In this example, we consider Model (2.1) with $\mathbf{f}_t$ being the same as that in Example 1. We set the true number of factors $r = 5$, the number of the spiked components $K = 3, 7$ defined in Assumption 5, the dimensions are $p = 50, 100, 300, 500$, and the sample sizes are $n = 300, 500, 1000, 1500, 3000$. We consider three scenarios for δ_1 and δ_2: $(\delta_1, \delta_2) = (0, 0)$, $(\delta_1, \delta_2) = (0.4, 0.5)$ and $(\delta_1, \delta_2) = (0.5, 0.4)$. For each setting, the elements of $\mathbf{L}$ are drawn independently from $U(-2, 2)$, and then we divide $\mathbf{L}_1$ by $p^{\delta_1/2}$, the first K columns of $\mathbf{L}_2$ by $p^{\delta_2/2}$ and the rest $v - K$ columns by p to satisfy Assumptions 4 and 5. $\mathbf{\Phi}$, $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are drawn similarly as those of Example 1. We use 1000 replications in each experiment.

We first study the performance of the high-dimensional white noise test. For simplicity, we only present the results of the $T(m)$ statistics defined in (2.12) and the results for the other test are similar. When $p \geq n$, we only keep the upper $0.75n$ components of $\widehat{\mathbf{G}}'\mathbf{y}_t$ in the testing. The results are reported in Table 2 for $r = 5, K = 3$ and Table 3 for $r = 5, K = 7$. From Tables 2 and 3, we see that for each setting of (δ_1, δ_2) and fixed p, the performance of the white noise test improves as the sample size increases. The performance is also quite satisfactory for moderately large p. In addition, the performance of the test when $\delta_1 < \delta_2$ is slightly better than that when $\delta_1 > \delta_2$ which is in agreement with Theorem 6 since the the convergence rate discussed in Remark 6 for Theorem 6(ii) is $p^{0.8}n^{-1/2}$ for the former and

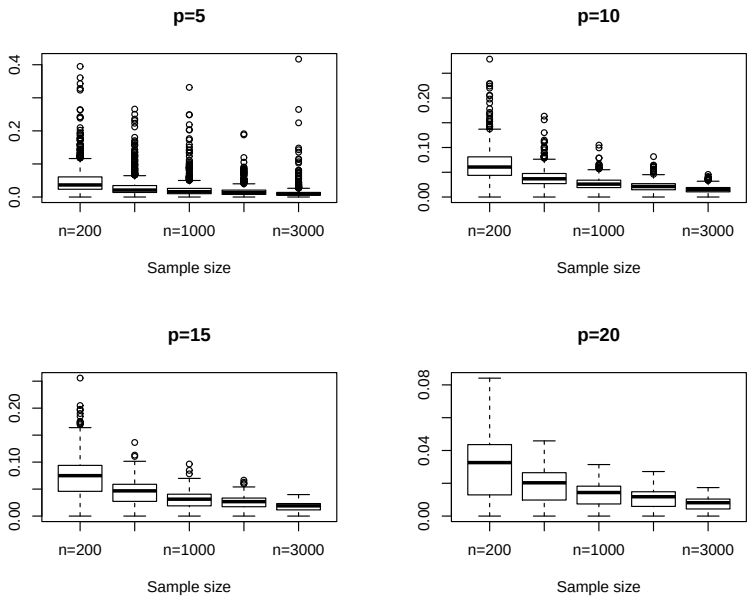

Figure 1: Boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{L}_1))$ when $r = 3$ under the scenario that p is relatively small in Example 1. The sample sizes are $200, 500, 1000, 1500, 3000$, respectively. 1000 iterations are used.

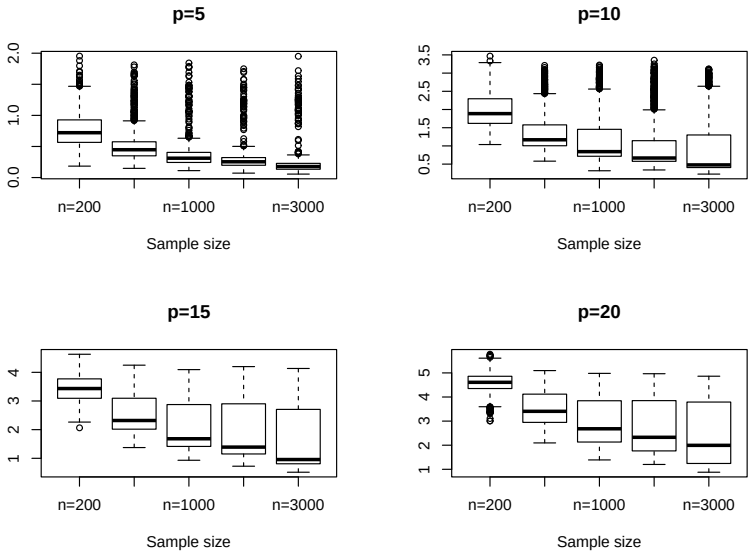

Figure 2: Boxplots of the RMSE defined in (4.2) when $r = 3$ under the scenario that p is relatively small in Example 1. The sample sizes are $200, 500, 1000, 1500, 3000$, respectively. 1000 iterations are used.

$p^{0.9}n^{-1/2}$ for the later one. Even though the performance of the test for $\delta_1 = \delta_2 = 0$ is better than that of $\delta_1 = 0.5$ and $\delta_2 = 0.4$ when the sample size is small (e.g. $n = 300$), but as the sample size increases, the performance of the test in the latter case is better than that in the former case because the convergence rate in the former case is pn^{-1}. This finding is consistent with the asymptotic theory in Theorem 6.

To shed some light on the advantages of the proposed methodology, we compare our method of selecting the number of factors with those in Bai and Ng (2002) and Lam et al. (2011). Specifically, for the principal components method in Bai and Ng (2002), the number of factors is determined by the BIC-type criterion, defined by

$$\widehat{r} = \arg\min_{1\le k\le \widetilde{k}} \left\{ \log(\frac{1}{np}\sum_{t=1}^{n} \|\widehat{\varepsilon}_t\|_2^2) + k\left(\frac{p+n}{np}\log(\frac{np}{p+n})\right)\right\}, \tag{4.3}$$

where we choose $\widetilde{k} = 20$ and $\widehat{\varepsilon}_t$ is the p-dimensional residuals obtained by the principal component analysis. For the ratio-based method in Lam et al. (2011), let $\widehat{\lambda}_1, ..., \widehat{\lambda}_p$ be the eigenvalues of $\widehat{\mathbf{M}}$, define

$$\widehat{r} = \arg\min_{1\le j\le R} \{\frac{\widehat{\lambda}_{j+1}}{\widehat{\lambda}_j}\}, \tag{4.4}$$

where we choose $R = p/2$ as suggested in their paper. Figures 3 and 4 present the boxplots of $\widehat{r}$. From Figure 3, we see that when the dimension is relatively small (e.g. $p = 50$), the criterion in (4.3) tends to overestimate the number of factors and it is far away from the true one. As p increases, when $r = 5, K = 3$, the estimated number of factors is $8(= r + K)$, which includes the number of factors and the spiked components in the noise covariance. The same issue occurs to the case when $r = 5$ and $K = 7$; see the lower panel in Figure 3. For the ratio based method, we see from Figure 4 that the estimated number of factors $\widehat{r}$ is the combination of the factors and the noise terms, which is similar to the BIC method for large p. Overall, we conclude that the information criterion and the ratio-based method may fail if the covariance matrix of the noise has diverging eigenvalues. On the other hand, the white noise test considered still works well.

Next, we study the accuracy of the estimated loading matrices as that in Example 1. The boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{L}_1))$ are shown in Figure 5. Similar pattern is also obtained for the estimation of other matrices, and we omit them here. From Figure 5, there is a clear pattern that the estimation accuracy of the loading matrix improves as the sample size increases even for moderately large p, which is in line with our asymptotic theory. The results also confirm

Table 2: Empirical probabilities $P(\widehat{r} = r)$ for Example 2 with $r = 5$ and $K = 3$, where p and n are the dimension and the sample size, respectively. δ_1 and δ_2 are the strength parameters corresponding to the factors and the errors, respectively. 1000 iterations are used.

		n				
(δ_1, δ_2)	p	300	500	1000	1500	3000
(0,0)	50	0.510	0.833	0.906	0.917	0.926
	100	0.538	0.799	0.910	0.916	0.922
	300	0.582	0.907	0.916	0.924	0.932
	500	0.560	0.888	0.918	0.928	0.932
(0.4,0.5)	50	0.717	0.903	0.928	0.929	0.935
	100	0.800	0.924	0.938	0.940	0.944
	300	0.858	0.904	0.928	0.932	0.952
	500	0.834	0.922	0.932	0.933	0.948
(0.5,0.4)	50	0.420	0.890	0.910	0.916	0.920
	100	0.508	0.868	0.912	0.928	0.936
	300	0.581	0.910	0.926	0.929	0.932
	500	0.678	0.928	0.936	0.938	0.934

Table 3: Empirical probabilities $P(\widehat{r} = r)$ of Example 2 with $r = 5$ and $K = 7$, where p and n are the dimension and the sample size, respectively. δ_1 and δ_2 are the strength parameters corresponding to the factors and the errors, respectively. 1000 iterations are used.

		n				
(δ_1, δ_2)	p	300	500	1000	1500	3000
(0,0)	50	0.418	0.688	0.904	0.908	0.910
	100	0.426	0.754	0.910	0.916	0.918
	300	0.406	0.686	0.914	0.925	0.926
	500	0.614	0.778	0.912	0.918	0.920
(0.4,0.5)	50	0.806	0.820	0.892	0.912	0.926
	100	0.800	0.914	0.922	0.904	0.922
	300	0.939	0.935	0.935	0.929	0.930
	500	0.898	0.904	0.926	0.930	0.933
(0.5,0.4)	50	0.332	0.856	0.900	0.928	0.938
	100	0.356	0.716	0.920	0.922	0.928
	300	0.384	0.688	0.924	0.936	0.945
	500	0.421	0.778	0.924	0.930	0.931

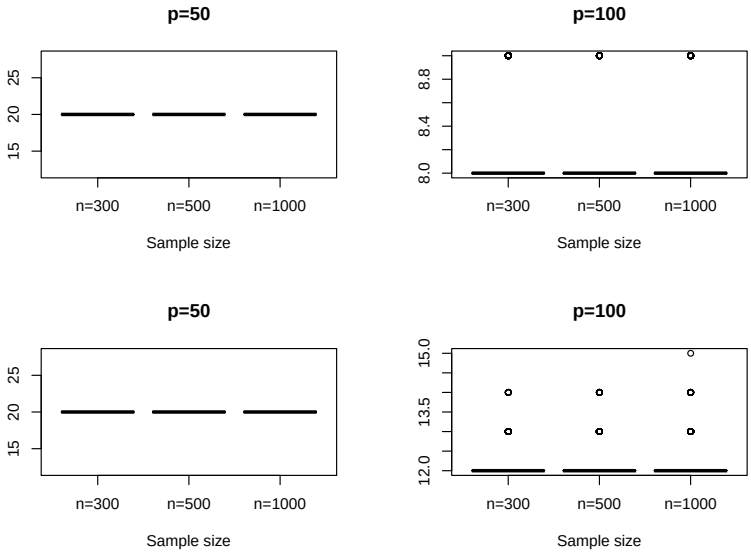

Figure 3: Boxplots of $\widehat{r}$ obtained by the information criterion method in (4.3) corresponding to BN when $r = 5$, $K = 3$ for the upper panel, and $K = 7$ for the lower panel of Example 2. 1000 iterations are used.

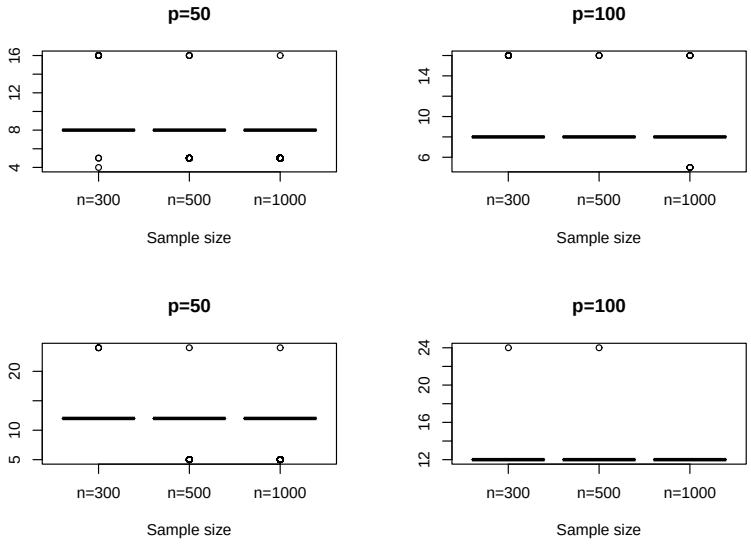

Figure 4: Boxplots of $\widehat{r}$ obtained by the ratio-based method in (4.4) corresponding to LYB when the true $r = 5$, $K = 3$ for the upper panel, and $K = 7$ for the lower panel of Example 2. 1000 iterations are used.

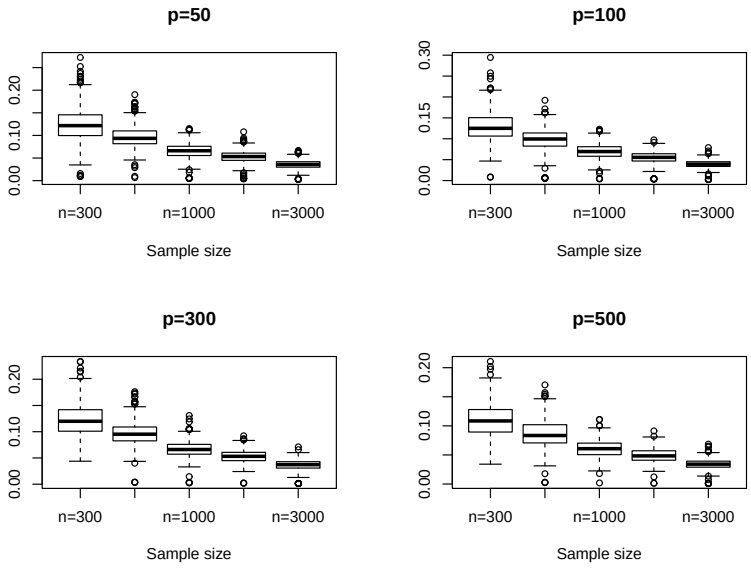

Figure 5: Boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{L}_1))$ when $r = 3$ and $K = 5$ under the scenario that p is relatively large in Examle 2. $n = 300, 500, 1000, 1500, 3000$, respectively. 1000 iterations are used.

that the proposed white noise test selects $\widehat{r}$ reasonably well even for large p.

Finally, we compare the proposed methodology with those in Bai and Ng (2002) and Lam et al. (2011) in terms of the RMSE defined below:

$$\text{RMSE} = \left(\frac{1}{np} \sum_{t=1}^{n} \|\widehat{\mathbf{A}}_1 \widehat{\mathbf{x}}_t - \mathbf{L}_1 \mathbf{f}_t\|_2^2 \right)^{1/2}, \tag{4.5}$$

which is different from that in Equation (4.2) since we have another factor $p^{-1/2}$ in (4.5). This RMSE quantifies the estimation accuracy of the common factor process. In the comparison, the number of factors are obtained by the corresponding methods of each methodology. The results are shown in Table 4 for $r = 5$, $K = 7$ and $\delta_1 = \delta_2 = 0$. The pattern is similar for the other settings. We denote **GT** as the proposed method, **LYB** as that in Lam et al. (2011), and **BN** as Bai and Ng (2002). When calculating $\widehat{\mathbf{B}}_2^*$ using our method, we choose the number of components $\widehat{K} = 10$, which is fixed in all the iterations. Thus, $\widehat{\mathbf{B}}_2^*$ contains $p - \widehat{K}$ columns corresponding to the $p - \widehat{K}$ smaller eigenvalues of $\widehat{\mathbf{S}}$. From the table, we see that, because the BIC and the ratio-based method tend to overestimate the number of common factors r in the presence of diverging eigenvalues in the covariance matrix of the idiosyncratic component, the RMSE of our method is much smaller than those obtained by Bai and Ng

Table 4: The RMSE defined in (4.5) when $r = 5$ and $K = 7$ in Example 2. $n = 300, 500, 1000, 1500, 3000$, respectively. Standard errors are given in the parentheses and 1000 iterations are used. GT denotes the proposed method, BN denotes the principal component analysis in Bai and Ng (2002) and LYB is the one in Lam et al. (2011)

		n				
Method	p	300	500	1000	1500	3000
GT		**1.510(0.233)**	**1.124(0.235)**	**0.770(0.235)**	**0.627(0.224)**	**0.488(0.273)**
LYB	50	3.056(0.085)	3.051(0.081)	3.056(0.075)	3.053(0.122)	2.976(0.400)
BN		3.058(0.086)	3.053(0.082)	3.058(0.075)	3.059(0.077)	3.055(0.074)
GT		**1.490(0.179)**	**1.148(0.188)**	**0.817(0.141)**	**0.677(0.126)**	**0.519(0.191)**
LYB	100	3.050(0.074)	3.056(0.065)	3.053(0.055)	3.046(0.159)	3.024(0.257)
BN		3.051(0.075)6	3.057(0.065)	3.054(0.055)	3.057(0.055)	3.052(0.052)
GT		**1.729(0.118)**	**1.463(0.107)**	**1.149(0.094)**	**1.107(0.079)**	**0.769(0.077)**
LYB	300	3.052(0.047)	3.055(0.047)	3.053(0.040)	3.056(0.037)	3.056(0.034)
BN		3.053(0.055)	3.056(0.047)	3.054(0.040)	3.056(0.037)	3.057(0.034)
GT		**1.753(0.089)**	**1.547(0.081)**	**1.285(0.052)**	**1.044(0.070)**	**0.861(0.047)**
LYB	500	3.057(0.053)	3.050(0.042)	3.054(0.035)	3.055(0.034)	3.055(0.027)
BN		3.058(0.053)	3.050(0.042)	3.054(0.035)	3.056(0.034)	3.055(0.027)

(2002) and Lam et al. (2011). Also, as expected, for a given p, the RMSE tends to decrease when the sample size increases. This is in agreement with the asymptotic theory in Theorem 5. Overall, under the reasonable assumption that the top eigenvalues of the noise covariance matrix are diverging for the high-dimensional case, the proposed method outperforms the existing ones in the literature.

4.2 Real data analysis

In this section, we apply the proposed method to a real example to illustrate its usefulness in practice. The dimension p is smaller than the sample size n in this example. An additional real example is shown in the online supplement, where the dimension p is greater than the sample size.

Example 3. In this example, we consider the daily returns of 49 Industry Portfolios which can be downloaded from `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. There are many missing values in the data so we only apply the proposed method to the period from July 13, 1988 to November 23, 1990 for a total of 600 observations. The series are shown in Figure 6, where we have $n = 600$ and $p = 49$. Applying the white noise test, we find that there are 6 common factors. In the testing, we use $k_0 = 5$ in Equation (2.9), $m = 10$ in the test statistic $T(m)$, and the upper 95%-quantile 2.97 of the Gumbel distribution as the critical value of the test. To recover the factors, we first examine

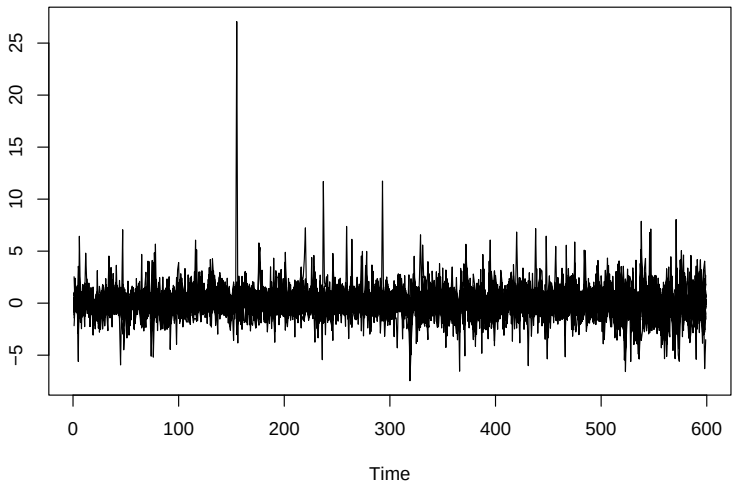

Figure 6: Time plots of daily returns of 49 Industry Portfolios with 600 observations from July 13, 1988 to November 23, 1990 of Example 3.

the eigenvalues of $\widehat{\mathbf{S}}$. Figure 7(a) shows the first 10 eigenvalues of $\widehat{\mathbf{S}}$ whereas Figure 7(b) plots the ratios of these eigenvalues. From the ratio-plot, we see that the largest drop of the ratios occurs at the first eigenvalue to the second. However, following the proposed procedure, we choose $\widehat{K} = \widehat{K}_L = \min\{\sqrt{p}, \sqrt{n}, 10\} = 7$ in our analysis. The spectral densities of the 6 estimated factors are shown in Figure 8. Note that the spectral densities hardly change if we vary $\widehat{K}$ from 1 to 10, but we do not report them here to save space. From the patterns of the spectral densities in Figure 8, we see that the estimated factors are all different from white noises. In this example, the largest eigenvalue of $\widehat{\mathbf{x}}_t$ is 10.74, which is almost at the same level as $\widehat{\mu}_1 = 7.14$ of $\widehat{\mathbf{S}}$ shown in Figure 7 with $p = 49$. This empirical phenomenon supports the assumption that the largest eigenvalue of the covariance matrix of the idiosyncratic terms tends to diverge for large p.

Next, we compare our method with those in Bai and Ng (2002) and Lam et al. (2011). First, for the principal component analysis, the estimated number of factors is $\widehat{r} = 11$ using the BIC in (4.3). The spectral densities of the first 9 estimated factors are shown in Figure 9. From the plots, we see that the 3rd and 5th extracted factors contain limited dynamic dependence because their spectral densities are flat. The two estimated factors are plotted in Figure 10 and the p-values of the Ljung-Box test statistic $Q(10)$ are 0.4016 for the 3rd and 0.1871 for the 5th factor, respectively. Therefore, these two estimated factors are essentially

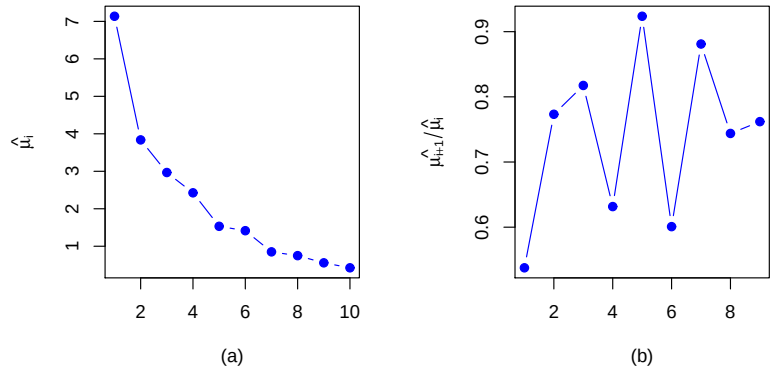

Figure 7: (a) The first 10 eigenvalues of $\widehat{\mathbf{S}}$ in Example 3; (b) The plots of the ratios for the eigenvalues $\widehat{\mu}_i$ of $\widehat{\mathbf{S}}$.

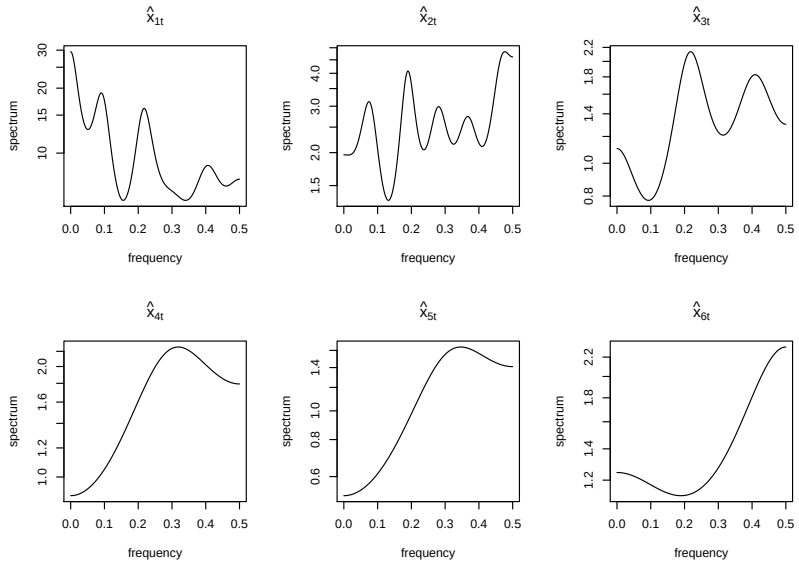

Figure 8: The spectral densities of 6 estimated common factors using the proposed methodology with $\widehat{K} = 7$ of Example 3.

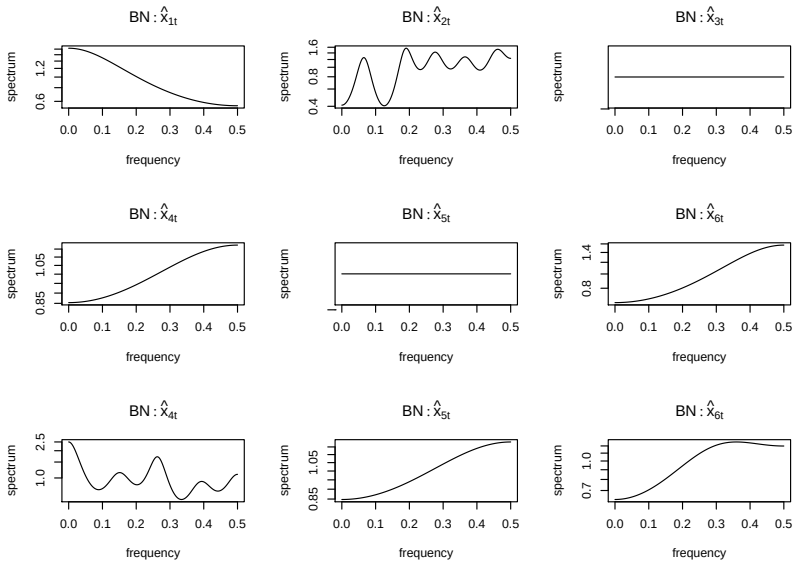

Figure 9: The spectral densities of the first 9 estimated factors using the principal component analysis in Bai and Ng (2002) of Example 3.

white noise processes. The results demonstrate a weakness of the principal component analysis, namely, it focuses on the covariance matrix of the data without paying any attention to the effect of the lagged variables. This weakness does not occur in the proposed method because it makes use of the relationships between the current and the lagged variables as shown in Figure 8.

For the ratio-based method in Lam et al. (2011), the estimated number of factors is $\widehat{r} = 1$. This phenomenon occurs often, implying that the method only picks the dominating signal. The spectral density functions of the first 6 transformed series $\widehat{u}_{1t}$,..., $\widehat{u}_{6t}$ are shown in Figure

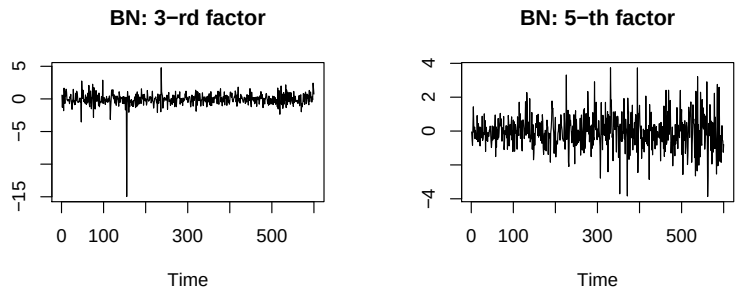

Figure 10: The time plots of the 3rd and 5th estimated factors using the principal component analysis in Bai and Ng (2002) of Example 3.

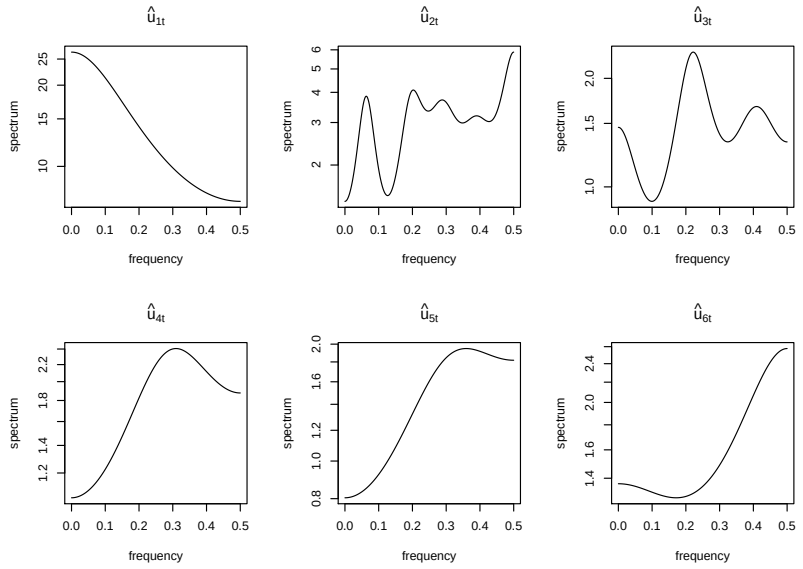

Figure 11: The spectral densities of first 6 transformed series using the eigen-analysis in Example 3.

11. Clearly, these 6 series are not white noise. Consequently, based on the assumptions of the procedure of Lam et al. (2011), these series are not idiosyncratic and should be included in the common factors. For financial returns, the first factor alone is not adequate in describing the behavior of the data, and many empirical analyses suggest that there are usually 3 or more factors affecting the financial market. See, for example, Fama and French (2015) and the references therein.

Finally, we compare the forecasting performance of the proposed method with those of other methods. For the h-step ahead forecasts, we compare the actual and predicted values of the model estimated using data in the time span $[1, \tau]$ for $\tau = 500, ..., 600 - h$, and the associated h-step ahead forecast error is defined as

$$\mathrm{FE}_h = \frac{1}{100 - h + 1} \sum_{\tau=500}^{600-h} \left(\frac{1}{\sqrt{p}} \|\widehat{\mathbf{y}}_{\tau+h} - \mathbf{y}_{\tau+h}\|_2 \right), \tag{4.6}$$

where $p = 49$ in this example. We first examine the estimated number of factors in the sub-samples using our method and those in Bai and Ng (2002) and Lam et al. (2011). The boxplots of the estimated $\widehat{r}$ for each τ are shown in Figure 12, and the means of $\widehat{r}$ obtained by the three methods are 6, 10.7 and 1, respectively. Therefore, we use $\widehat{r} = 6, 11$ and 1, respectively, for each τ. In addition, we employ VAR(1)-VAR(3) models to fit the

factor processes obtained by our method and the principal component analysis in Bai and Ng (2002), and scalar AR(1)-AR(3) models to fit the single factor process obtained by the ratio-based method in Lam et al. (2011). For simplicity, we use AR to denote AR model for a univariate process or VAR models for a multivariate process. The h-step ahead forecast errors are reported in Table 5 for $h = 1, 2, 3$, and similar patterns can be found for other forecast horizon h. In Table 5, we vary $\widehat{K}$ from 1 to 7 and the values in boldface represent the smallest ones using AR(1) to AR(3) models, respectively. From the table, we see that for the 1-step ahead forecasts the performance of the proposed method is slightly worse than that of Bai and Ng (2002), but the proposed method fares better than either BN or LYB method for 2-step and 3-step ahead forecasts, especially for $\widehat{K} \geq 5$. The result is understandable because the PCA method in Bai and Ng (2002) extracts the most significant coordinates of the data and, hence, it might produce more precise forecasts in the short term, but the increased variability associated with longer horizon is likely to decrease the accuracy in forecasting. On the other hand, the proposed method seems to be more stable in the forecasting performance, especially for the longer forecast horizon h. As an illustration, the point-wise forecast errors of the 1-step ahead prediction using AR(1) models and $\widehat{K} = 1$ are shown in Figure 13, where we also choose a random walk as the benchmark procedure. From Figure 13, we see that the three methods perform rather similarly and there are times our method produces smaller errors. We also note that all methods are better than the benchmark. In practice, we may find an optimal $\widehat{K}$ based on some cross-validation if the main interest of data analysis is prediction.

In this application, the factors identified by the proposed method appears to be reasonable, and they also fare well in out-of-sample forecasts. The principal component analysis of Bai and Ng (2002) extracts the components with large variances while overlooks the dynamical dependence in the data. In fact, the estimated factors themselves may be white noise. The ratio-based method of LYB only extracts the component associated with the largest eigenvalue, which may not be sufficient and makes model interpretation hard.

5 Discussion and Concluding Remarks

This article introduced a new structured factor model for high-dimensional time series analysis. We allow the largest eigenvalues of the covariance matrix of the idiosyncratic components

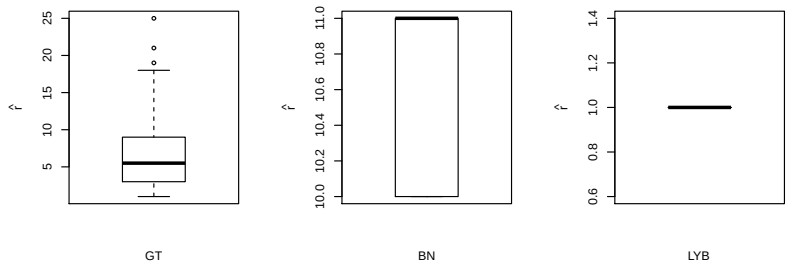

Figure 12: The boxplots of the estimated number of factors via various methods considered in the paper using the return data of Example 3 in the time span $[0, \tau]$ for $\tau = 500, ..., 599$. GT denotes the proposed method, BN denotes the principal component analysis in Bai and Ng (2002) and LYB is the one in Lam et al. (2011).

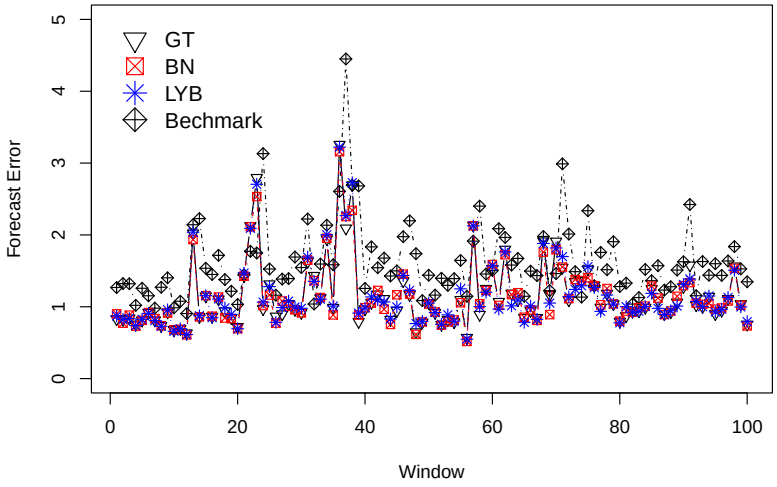

Figure 13: Time plots of the 1-step ahead point-wise forecast errors using AR(1) and VAR(1) models with $\widehat{K} = 1$ for various methods used in Example 3.

Table 5: The 1-step, 2-step and 3-step ahead forecast errors. Standard errors are given in the parentheses. GT denotes our method, BN denotes the principal component analysis in Bai and Ng (2002) and LYB is the one in Lam et al. (2011). Boldface numbers denote the smallest one for a given model.

		GT							BN	LYB
		$\widehat{K}=1$	$\widehat{K}=2$	$\widehat{K}=3$	$\widehat{K}=4$	$\widehat{K}=5$	$\widehat{K}=6$	$\widehat{K}=7$		
1-step	AR(1)	1.152	1.161	1.159	1.162	1.158	1.158	1.159	**1.142**	1.157
		(0.469)	(0.484)	(0.482)	(0.489)	(0.487)	(0.483)	(0.487)	(0.442)	(0.465)
	AR(2)	1.164	1.165	1.166	1.168	1.164	1.165	1.164	1.156	1.162
		(0.474)	(0.480)	(0.482)	(0.493)	(0.486)	(0.483)	(0.485)	(0.446)	(0.470)
	AR(3)	1.170	1.172	1.172	1.174	1.169	1.170	1.168	1.168	1.162
		(0.477)	(0.485)	(0.489)	(0.498)	(0.493)	(0.493)	(0.496)	(0.441)	(0.470)
2-step	AR(1)	1.179	1.180	1.180	1.180	1.179	**1.178**	**1.178**	1.182	1.180
		(0.512)	(0.512)	(0.512)	(0.513)	(0.512)	(0.510)	(0.510)	(0.513)	(0.514)
	AR(2)	1.190	1.190	1.190	1.188	1.188	1.187	1.185	1.197	1.185
		(0.519)	(0.514)	(0.514)	(0.513)	(0.514)	(0.512)	(0.512)	(0.520)	(0.519)
	AR(3)	1.194	1.193	1.194	1.191	1.191	1.191	1.189	1.204	1.185
		(0.520)	(0.519)	(0.520)	(0.519)	(0.520)	(0.520)	(0.523)	(0.510)	(0.520)
3-step	AR(1)	1.181	**1.180**	**1.180**	**1.180**	**1.180**	**1.180**	**1.180**	1.184	1.184
		(0.511)	(0.511)	(0.511)	(0.510)	(0.511)	(0.510)	(0.510)	(0.514)	(0.513)
	AR(2)	1.185	1.183	1.183	1.183	1.183	1.182	1.182	1.190	1.187
		(0.510)	(0.510)	(0.508)	(0.508)	(0.508)	(0.507)	(0.508)	(0.514)	(0.512)
	AR(3)	1.187	1.184	1.184	1.184	1.184	1.184	1.184	1.198	1.188
		(0.517)	(0.513)	(0.513)	(0.512)	(0.514)	(0.518)	(0.520)	(0.510)	(0.514)

to diverge to infinity by imposing some structure on the noise terms. The first step of the proposed analysis is an eigen-analysis of the matrix $\widehat{\mathbf{M}}$ defined in Equation (2.9). The form of $\widehat{\mathbf{M}}$ is a special case of the orthonormalized Partial Least Squares on time series data by assuming the covariance matrix of the data is identity. By an abuse of notation, let $\mathbf{w}_t = (\mathbf{y}'_{t-1}, ..., \mathbf{y}'_{t-k_0})'$ be the vector of past k_0 lagged values of the time series $\mathbf{y}_t$, where k_0 is a pre-specified positive integer as that in (2.6) and (2.9). The orthonormalized Partial Least Squares computes the orthogonal score vectors for $\mathbf{y}_t$ by solving the following optimization problem:

$$\max_{\mathbf{a}_i} \|E(\mathbf{a}'_i \mathbf{y}_t \mathbf{w}'_t)\|_2^2, \quad \text{subject to} \quad \mathbf{a}'_i E(\mathbf{y}_t \mathbf{y}'_t)\mathbf{a}_i = 1. \tag{5.1}$$

See, for example, Arenas-García and Camps-Valls (2008). It can be shown that the columns $\mathbf{a}_i$ are given by the principal eigenvectors of the following generalized eigenvalue problem:

$$\boldsymbol{\Sigma}_{yw}\boldsymbol{\Sigma}'_{yw}\mathbf{a}_i = \eta \boldsymbol{\Sigma}_y \mathbf{a}_i. \tag{5.2}$$

Note that $\mathbf{M} = \boldsymbol{\Sigma}_{yw}\boldsymbol{\Sigma}'_{yw}$ which is just the form in (2.6). To solve the above equation, we need to obtain accurate estimates for the covariance matrix and its inverse simultaneously, which however is not easy. Instead we change the subject condition in (5.1) to $\mathbf{a}'_i\mathbf{a}_i = 1$ and apply the eigen-analysis on $\widehat{\mathbf{M}}$ in (2.9), and this approach remains an effective way if we assume

the component variances of the data are uniformly bounded. In this case, the second step is needed.

The second step of the proposed analysis is the projected PCA on $\widehat{\mathbf{S}}$ in (2.10) by assuming the largest K eigenvalues of the covariance matrix of the idiosyncratic component are diverging. In practice, the most useful assumption is that the largest eigenvalue is diverging whereas the rests are bounded. Limited experience indicates that many real datasets share such a phenomenon. If we are only concerned with the forecasting performance of the proposed analysis, we may select $\widehat{K}$ in a range such as $\widehat{K}_L \leq \widehat{K} \leq \widehat{K}_U$, where $\widehat{K}_L$ and $\widehat{K}_U$ are defined in Remark 5(ii), via some cross-validation method like out-of-sample testing.

The white noise test considered is an efficient way to determine the number of common factors. The one-by-one bottom-up testing procedure may not perform well when the dimension p is high, but the limiting distribution of the test statistic of Tsay (2018) holds for large p by making use of the limiting theorems for the extreme value theory. If we like to use the test statistic for a wide range of dimensions and various sample sizes, we may adopt the small-sample adjustments for the test statistic discussed by the author. The simulation results in Tsay (2018) show that the resulting test statistic works reasonably well.

In conclusion, the proposed model and approach are natural and useful in analyzing high-dimensional time series data. The produced factors are meaningful and interpretable, and the forecast performance of the proposed method is as good as the principal component analysis and the ratio-based method commonly used in the literature.

Supplementary Material

The supplementary material contains all technical proofs of the theorems in Section 3 and an additional real example consisting of half-hourly temperature data observed at the Adelaide Airport in Australia with $p = 508$ and $n = 336$.

References

Arenas-García, J., and Camps-Valls, G. (2008). Efficient kernel orthonormalized PLS for remote sensing applications. *IEEE Transactions on Geoscience and Remote Sensing*, **46(10)**, 2872–2881.

Bai J. (2003) Inferential theory for factor models of large dimensions. *Econometrica*, **71(1)**, 135–171.

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

Black, F. (1986). Noise. *The Journal of Finance*, **41(3)**, 528–543.

Chang, J., Yao, Q., and Zhou, W. (2017). Testing for high-dimensional white noise using maximum cross-correlations. *Biometrika*, **104(1)**, 111–127.

Davis, R. A., Zang, P., and Zheng, T. (2012). Sparse vector autoregressive modelling. Available at *arXiv:1207.0520*.

Fama, E. F., and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, **116(1)**, 1–22.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). Reference cycles: the NBER methodology revisited (No. 2400). Centre for Economic Policy Research.

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

Gao, Z., Ma, Y., Wang, H., and Yao, Q. (2018). Banded spatio-temporal autoregressions. *Journal of Econometrics:* To appear.

Gao, Z., and Tsay, R. S. (2018). A structural-factor approach for modeling high-dimensional time series *Manuscript,* University of Chicago.

Gregory, A. W., and Head, A. C. (1999). Common and country-specific fluctuations in productivity, investment, and the current account. *Journal of Monetary Economics*, **44(3)**, 423–451.

Lam, C., and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, **40(2)**, 694–726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, **98**, 901–918.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis,* Springer, Berlin.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, **92(4)**, 1004–1016.

Pan, J., and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, **95(2)**, 365–379.

Shen, D., Shen, H., and Marron, J. S. (2016). A general framework for consistency of principal component analysis. *Journal of Machine Learning Research*, **17(150)**, 1–34.

Shojaie, A., and Michailidis, G. (2010). Discovering graphical Granger causality using the truncated lasso penalty. *Bioinformatics*, **26**, 517–523.

Song, S., and Bickel, P. J. (2011). Large vector auto regressions. Available at *arXiv:1106.3519*.

Stock, J. H., and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, **4**, 351–394.

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

Stock, J. H., and Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. Available at *www.nber.org/papers/w11467*.

Tiao, G. C., and Tsay, R. S. (1989). Model specification in multivariate time series (with discussion). *Journal of the Royal Statistical Society,* **B51**, 157–213.

Tsay, R. S. (2014). *Multivariate Time Series Analysis*. Wiley, Hoboken, NJ.

Tsay, R. S. (2018). Testing for serial correlations in high-dimensional time series via extreme value theory. *Manuscript,* University of Chicago.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press.