Bernoulli **28**(2), 2022, 1224–1249
https://doi.org/10.3150/21-BEJ1384

Non-asymptotic properties of spectral decomposition of large Gram-type matrices and applications

LYUOU ZHANG[1], WEN ZHOU[2,*] and HAONAN WANG[2,†]

[1]*School of Statistics and Management, Shanghai University of Finance of Economics, 777 Guoding Road, Shanghai, 200433, P.R. China. E-mail: zhanglvou@mail.shufe.edu.cn*
[2]*Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.*
*E-mail: *riczw@stat.colostate.edu; †wanghn@stat.colostate.edu*

Gram-type matrices and their spectral decomposition are of central importance for numerous problems in statistics, applied mathematics, physics, and machine learning. In this paper, we carefully study the non-asymptotic properties of spectral decomposition of large Gram-type matrices when data are not necessarily independent. Specifically, we derive the exponential tail bounds for the deviation between eigenvectors of the right Gram matrix to their population counterparts as well as the Berry-Esseen type bound for these deviations. We also obtain the non-asymptotic tail bound of the ratio between eigenvalues of the left Gram matrix, namely the sample covariance matrix, and their population counterparts regardless of the size of the data matrix. The documented non-asymptotic properties are further demonstrated in a suite of applications, including the non-asymptotic characterization of the estimated number of latent factors in factor models and relate machine learning problems, the estimation and forecasting of high-dimensional time series, the spectral properties of large sample covariance matrix such as perturbation bounds and inference on the spectral projectors, and low-rank matrix denoising using dependent data.

Keywords: Approximate factor model; Gram-type matrices; high-dimensional time series; non-asymptotic analysis; principal component analysis; spectral decomposition

1. Introduction

Gram-type matrix or Gram matrix is fundamental in a wide range of fields including statistics (Shawe-Taylor et al. [83]), applied mathematics (James and Murphy [58], Schölkopf et al. [81], Shawe-Taylor et al. [82], Chen, Womersley, and Ye [39]), machine learning (Drineas and Mahoney [45], De Almeida, Asada, and Garcia [42], Ramona, Richard, and David [79]), engineering (De Almeida, Asada, and Garcia [43]), and physics (Stark [84]). Given a $p \times T$ data matrix $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ with p-dimensional observation $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{pt})^\top$, the *left* and the *right* Gram matrices are $\mathbf{Y}\mathbf{Y}^\top$ and $\mathbf{Y}^\top\mathbf{Y}$, respectively (Horst [56], Rummel [80]). Statistically, the left Gram matrix scaled by the sample size $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ coincides with the sample covariance matrix after ignoring the sample mean. As a bilinear function of the data matrix, Gram matrix retains important information about data. For example, the right Gram matrix and the data matrix share the common null space while the column space of the left Gram matrix agrees with that of the data matrix. Particularly, the spectral decomposition of Gram matrices is a powerful and popular tool to provide a low-rank representation of the original data yet preserves the information as much as possible. For instance, in the linear model, spectral decomposition of the Gram matrix from the design matrix reveals the direction of space spanned by the projection matrix (Mandel [71]); in the nonparametric regression, spectral decomposition of the Gram matrix from the spline basis functions provides a complete reconstruction of the functional space (Bialecki and Fairweather [21]); and in the exploratory analysis, spectral decomposition of the Gram matrix from a general data or

feature matrix leads to the principal component analysis (PCA) (Pearson [78], Hotelling [57], Jolliffe [65]), kernel PCA, or sparse PCA (Zou, Hastie, and Tibshirani [95], Zou and Xue [96]). In addition, spectral decomposition of the Gram matrix has been applied to estimate large covariance matrices (Fan, Liao, and Micheva [47], Fan et al. [49]) and to extract the latent factors that drive the correlation structure in factor models (Bai [8], Bartholomew, Knott, and Moustaki [20], Bai and Ng [13], Fan, Liao, and Wang [48]). By itself, the spectral decomposition has also been applied to other type of matrices to reveal the underlying structure in data, such as the spectral method along with the graph Laplacian or the adjacency matrix in cluster analysis or network study for the detection of clusters or latent communities (Donath and Hoffman [44], Ng, Jordan, and Weiss [75]).

Gram matrix naturally grows along with the size of data, and not only it may incur computational challenges but also lead to theoretical difficulties. For fixed dimensions, the scaled left Gram matrix or the sample covariance matrix converges to its expectation when T diverges (Bai, Yin, and Krishnaiah [17,18], Bai and Yin [16]). However, both the left and the right Gram matrices, as well as their empirical spectral distributions may fail to converge given simultaneously divergent p and T (Bickel and Levina [22,23], Johnstone and Lu [63], Wang and Fan [88]). Based on the asymptotic normality of sample covariance matrix, Anderson [5] established the joint distribution of empirical eigenvalues in the asymptotic regime where p remains constant and T diverges. For independent and identically distributed (*i.i.d.*) data with divergent dimensions, which scale with the sample size linearly and vice versa, the limiting distribution of spectral structures of the sample covariance matrix has also been widely studied (Wachter [87], Jonsson [66], Bai, Yin, and Krishnaiah [17,18], Bai and Yin [16], Adamczak et al. [1], Bai and Silverstein [15]). When p/T diverges, a flexible and common approach is the spike structure model (Johnstone [62]). That is, among the p eigenvalues of the population covariance matrix of $\boldsymbol{y}_t$, there are K dominant eigenvalues compared to the remains so that the signal of low-rank structure outweighs the noise and therefore can be retrieved from the spectral decomposition. Leveraging this spike structure, Wang and Fan [88] showed that, for divergent p/T, the eigenvalue and corresponding eigenvector of the sample covariance matrix still converge to their population counterparts whenever the K dominant population eigenvalues diverge in p with certain rate. They also showed that the convergence rates of empirical eigenvalue and eigenvector are controlled by the divergent rate of the corresponding population eigenvalue.

The aforementioned assumption that the first K dominant eigenvalues of the population covariance matrix of $\boldsymbol{y}_t$ have order $O(p)$, together with the assumption that noises admit constant variance, is known as the *pervasiveness assumption* or *strong factor assumption* from the factor model and econometrics literature. Under this assumption, the spike structure can be equivalently written as a factor model (Chamberlain and Rothschild [34], Stock and Watson [86], Bai [8], Lam and Yao [69]) for which data satisfies

$$y_{it} = a_{i1} f_{t1} + \cdots + a_{iK} f_{tK} + u_{it} \tag{1.1}$$

with $t = 1, \ldots, T$ and $i = 1, \ldots, p$. Here, $(f_{t1}, \ldots, f_{tK})^\top$ is a K-dimensional zero mean latent process and u_{it} is an error process. Model (1.1) inherently links to a large number of widely used statistical models and methods, such as the panel data model with unobservable interactive effects (Ahn, Lee, and Schmidt [3], Bai [9], Bai and Li [10], Moon and Weidner [73]) and PCA (Fan et al. [49]). In matrix form, (1.1) is

$$\mathbf{Y} = \mathbf{A}\mathbf{F}^\top + \mathbf{U}, \tag{1.2}$$

where $\mathbf{A} = (a_{ik})_{i=1,k=1}^{p,K}$, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)^\top$ with $\mathbf{f}_t = (f_{t1}, \ldots, f_{tK})^\top$ or equivalently, $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K)$ with $\boldsymbol{f}_k = (f_{1k}, \ldots, f_{Tk})^\top$, and $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ with $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{pt})^\top$. Assume that $\mathbf{f}_t$ and

$\boldsymbol{u}_t$ are uncorrelated and $\mathbb{E}(\mathbf{f}_t\mathbf{f}_t^\top) = \mathbf{I}_K$ for each $t = 1, \ldots, T$ (Chamberlain and Rothschild [34]), the covariance of $\boldsymbol{y}_t$ is then given by

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Sigma}_u, \tag{1.3}$$

where $\boldsymbol{\Sigma} = T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$ and $\boldsymbol{\Sigma}_u = T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$. Model (1.2) is called the strict factor model if $T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ is diagonal, *i.e.*, $u_{1t}, \ldots, u_{pt}$ are uncorrelated with each other; otherwise, it is called the approximate factor model if $T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ is not diagonal (Chamberlain and Rothschild [34]). Model (1.2) provides an effective dimension reduction by approximating a p-dimensional process $\boldsymbol{y}_t$ with a K-dimensional process $\mathbf{f}_t$ and a loading matrix matrix $\mathbf{A}$. From (1.3), it is easy to see that the largest K eigenvalues of $\boldsymbol{\Sigma}$ increase in p while the remaining eigenvalues are bounded (Bai and Ng [12]), which mimics the spike structure model with divergent spiked eigenvalues.

For the traditional factor model with fixed p and *i.i.d.* normally distributed $\mathbf{f}_t$ and $\boldsymbol{u}_t$, the column space of loading matrix $\mathbf{A}$ and the diagonal entries of $T^{-1}\mathbb{E}(\mathbf{U}\mathbf{U}^\top)$ can be consistently estimated through either the maximum likelihood estimator (MLE) (Lawley and Maxwell [70]) or PCA (Anderson and Rubin [6], Anderson [4]), both of which rely on the consistent estimation of $\boldsymbol{\Sigma}$. Though factor models and PCA are not identical in general, they are approximately the same for high-dimensional problems under the pervasiveness assumption (Fan, Liao, and Micheva [47], Fan et al. [49]). Specially, the principal components $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ are defined as $\mathbf{Z}_k = \boldsymbol{w}_k^\top\mathbf{Y}$, where the projection directions $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K \in \mathbb{R}^p$ are the first K eigenvectors of $\boldsymbol{\Sigma}$. This eigen-decomposition formulation of PCA relates PCA to the singular value decomposition (SVD) of $\mathbf{Y}$ as well as the spectral decomposition of the sample covariance matrix, namely the left Gram matrix of $\mathbf{Y}$ scaled by sample size T.

In this paper, to carefully study the spectral decomposition of large Gram matrices, we consider data generated from (1.1) or (1.2) so that not only the data are of high-dimensional but also allow temporally dependence. For the right Gram matrix $\mathbf{Y}^\top\mathbf{Y}$, the eigenvectors corresponding to the K largest eigenvalues are of the same direction as $\boldsymbol{f}_k$, where $\boldsymbol{f}_k$ is the kth column of $\mathbf{F}$. Therefore, the spectral decomposition of the right Gram matrix can be investigated using the estimates to latent factor process and loading matrix in (1.1). That is, given an estimator to $\boldsymbol{f}_k$, denoted by $\widehat{\boldsymbol{f}}_k$, properties of the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}^\top\mathbf{Y}$ can be studied from $T^{-1/2}\widehat{\boldsymbol{f}}_k$, and vice versa. Although the consistency of estimating $\mathbf{F}$ has been documented in literature (Bai and Ng [13], Fan, Liao, and Wang [48]), non-asymptotic properties of the deviation of $\widehat{\mathbf{F}} = (\widehat{\boldsymbol{f}}_1, \ldots, \widehat{\boldsymbol{f}}_K)$, where $\widehat{\boldsymbol{f}}_k$ is the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{Y}^\top\mathbf{Y}$, from $\mathbf{F}$ have not been fully investigated. Our *main contribution* in this paper is to study the non-asymptotic properties of $\widehat{\mathbf{F}} - \mathbf{F}$ as well as the approximated distribution of $\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k$ for each k. Particularly, we relax the condition on $\mathbf{F}$ in the traditional factor model. Compared with Condition PC1 in Bai and Ng [13], we do not restrict $\mathbf{F}$ on a subspace. Therefore, as an important application in modeling high-dimensional time series, the non-asymptotic characterization of $\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k$ shows the accuracy of $\widehat{\boldsymbol{f}}_k$ as an surrogate to $\boldsymbol{f}_k$ for each k so that the parametric model of the K-dimensional latent processes, if specified in advance, can be easily estimated and therefore can be employed to forecast $\boldsymbol{y}_t$. Compared to the traditional likelihood based approach, this approach is computationally easier and requires very little assumptions on innovations of processes. In addition, we obtain non-asymptotic properties of the deviation between eigenvectors corresponding to the largest K eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$, *i.e.*, the sample covariance matrix, to those of $\boldsymbol{\Sigma}$ in (1.3). By considering $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ as a perturbation of $\boldsymbol{\Sigma}$, our result is similar to the Davis-Kahan Theorem (Davis and Kahan [41], Yu, Wang, and Samworth [90], Fan, Wang, and Zhong [50], Zhang, Cai, and Wu [91]) or the Wedin Theorem (Wedin [89]). Our conclusion, however, does not depend on the consistent estimation of $\boldsymbol{\Sigma}$. Hence, for the high-dimensional cases, our result remains valid for the spike part of $\boldsymbol{\Sigma}$ even though it cannot be consistently estimated using $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ without regularization.

Another important application of our results is to provide the non-asymptotic characterization of the tail probability of correctly estimating the number of latent factors K in the factor models, without

which recovering the latent factor processes and their loadings will be meaningless in practice. For fixed or low dimensions, a variety of subjective methods such as scree plot of eigenvalues, distribution-based tests including Bartlett's test, and computational intensive methods including cross-validation have been employed to determine K (Jolliffe [65]). For high dimensions with p/T converging to some constant, the information criteria such as AIC and BIC has been employed (Bai and Ng [11], Bai, Choi, and Fujikoshi [14]). If the data also follows a normal distribution, a sequential Kac-Rice test has been introduced to select K (Choi, Taylor, Tibshirani [40]). For ultra high dimensions with $p \gg T$, from the fact that the largest K eigenvalues of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ grow rapidly in p while others remain bounded or grow much slower, the consecutive-eigenvalue type estimator is widely used to determine K. For example, Lam and Yao [69] and Ahn and Horenstein [2] proposed estimators of K based on the ratios of consecutive eigenvalues. A similar approach is to use the difference of consecutive eigenvalues (Onatski [76]). These early results focus on the consistency of the estimated number of factors when p and T diverge. To better understand how the dimension and sample size affect the probability of correctly estimating the number of latent factors using those consecutive-eigenvalue type estimators, we first refine results regarding eigenvalues of the sample covariance matrix (Bai and Yin [16], Johnstone [62]). Then, we obtain non-asymptotic properties of the ratio of consecutive eigenvalues of the sample covariance matrix, which further provides the desired exponential tail bound of the probability of correctly estimating K for factor models or related machine learning problems.

The paper is organized as follows. In Section 2, we collect the notation and discuss the preliminary conditions to derive the main results. In Section 3, we carry out a non-asymptotic analysis of the spectral decomposition of large Gram matrices and document the main results. In Section 4, we discuss a variety of applications of our results to high-dimensional statistics. Section 5 presents numerical studies to demonstrate our results in the applications. We conclude the paper in Section 6 and relegate all the proofs and technical details to the Supplementary Materials (Zhang, Zhou, and Wang [93]).

2. Notation and preliminary conditions

We collect notation in Section 2.1 that will be used throughout the paper and discuss in detail the preliminary assumptions in Section 2.2 to establish the main results.

2.1. Notation

For p-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, its ℓ_q-norm is defined by $||\boldsymbol{a}||_q = (\sum_{j=1}^p |a_j|^q)^{1/q}$ with $1 \le q < \infty$. For matrix $\mathbf{M} = (m_{ij})_{1\le i,j\le p} \in \mathbb{R}^{p\times p}$, $\|\mathbf{M}\|_{\max} = \max_{i,j} |m_{ij}|$ denotes the maximum norm and $||\mathbf{M}||_{\mathbb{F}} = (\sum_{i=1}^p \sum_{j=1}^p m_{ij}^2)^{1/2}$ is the Frobenius norm. The spectral norm of $\mathbf{M}$ corresponds to its largest singular value, defined as $||\mathbf{M}||_2 = \sup_{\boldsymbol{a}\in S} ||\mathbf{M}\boldsymbol{a}||_2$, where $S = \{\boldsymbol{a} \in \mathbb{R}^p : ||\boldsymbol{a}||_2 = 1\}$. Denote the minimum and maximum eigenvalues of $\mathbf{M}$ by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$, respectively. Let $\mathrm{tr}(\mathbf{M}) = \sum_{j=1}^p m_{jj}$ be the trace of $\mathbf{M}$. For sequences $\{a_n\}$ and $\{b_n\}$, $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$ and $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n|/b_n < \infty$; $X_n = o_p(a_n)$ and $X_n = O_p(a_n)$ are similarly defined for a sequence of random variables X_n; $a_n \lesssim b_n$ if and only if $a_n \le Cb_n$ for some positive C independent of n; and $a_n \asymp b_n$ if and only if there exist positive constants C and D independent of n such that $Cb_n \le a_n \le Db_n$. Unless specified otherwise, $s > 1$ and $C > 0$ denote generic constants independent of p, T.

2.2. Conditions

Suppose one observes data $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{it}, \ldots, y_{pt})$ from model (1.1) or (1.2) with $t = 1, \ldots, T$. We pose the following conditions throughout the paper.

Condition 2.1. Almost surely, $\mathbf{A}^\top \mathbf{A}$ is a diagonal matrix with distinct entries; for each t, $f_{t1}, \ldots, f_{tK}$ are uncorrelated with each other and have zero mean and unit variance; for each t, $u_{1t}, \ldots, u_{pt}$ have zero mean and finite variances; and $\mathbf{f}_t$ and $\boldsymbol{u}_t$ are independent with each other.

Condition 2.1 is similar to the assumption imposed on the approximate factor model (Chamberlain and Rothschild [34]), which leads to the decomposition and identification of $\boldsymbol{\Sigma}$ in (1.3). The assumption on $\mathbf{A}$ can be viewed as Condition PC1 for the traditional factor models (Bai and Ng [13]), which is also imposed for the MLE by Lawley and Maxwell [70].

Condition 2.2. There exist constants $d_1, d_2 > 0$ such that $d_1 \le \lambda_{\min}\left(p^{-1}\mathbf{A}^\top \mathbf{A}\right) \le \lambda_{\max}\left(p^{-1}\mathbf{A}^\top \mathbf{A}\right) \le d_2$.

Since the largest K eigenvalues of $\mathbf{A}^\top \mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are the same, the spiked eigenvalues of $\boldsymbol{\Sigma}$ essentially diverge at rate p under Condition 2.2. When the entries of $\mathbf{A}$ remain constants as p diverges, this is always satisfied for a full rank $\mathbf{A}$ under Condition 2.1. In general, Condition 2.2 implies that, for each $k = 1, \ldots, K$, the mean squared loadings of the kth factor satisfies $p^{-1}\sum_{i=1}^p a_{ik}^2 = O(1)$, which can be easily satisfied with high probability if a_{ik} are *i.i.d.* copies from some non-degenerate distribution.

Condition 2.3. Denote $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ the σ-algebra generated by $\{(\mathbf{f}_t, \boldsymbol{u}_t) : t \le 0\}$ and $\{(\mathbf{f}_t, \boldsymbol{u}_t) : t \ge T\}$, respectively; define the mixing coefficient $\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A \cap B)|$; and denote $\#\{\mathcal{S}\}$ the cardinality of the set $\mathcal{S}$.

(i) Stationarity: $\{\boldsymbol{u}_t, \mathbf{f}_t\}_{t \le T}$ are weakly stationary.
(ii) Strong mixing across t: there exist $r_1, C_1 > 0$ such that $\alpha(s) < \exp(-C_1 s^{r_1})$ for any $s > 0$.
(iii) Weak dependence in errors: there exist $\gamma > 1/2$ and $C_2 > 0$ such that $\max_{1 \le j \le p} \sum_{i=1}^p |\mathbb{E}(u_{it}u_{jt})| < C_2$, $(pT)^{-1}\sum_{i=1}^p \sum_{j=1}^p \sum_{t=1}^T \sum_{s=1}^T |\mathbb{E}(u_{it}u_{js})| < C_2$; $\max_{1 \le i,i' \le p} \#\{(k,m) : p^{-2-\gamma} < S_1(i,i') < (p^\gamma \log p)^{-1}\} < C_2 \log p$, $\max_{1 \le i,i' \le p} \#\{(k,m) : S_1(i,i') > (p^\gamma \log p)^{-1}\} < C_2$, where $S_1(i,i',k,m) = T^{-2}\sum_{t=1}^T \sum_{s=1}^T |\operatorname{Cov}(u_{it}u_{kt}, u_{i's}u_{ms})|$, and; $\max_{1 \le i,i' \le p} \#\{(k,k',m,m') : p^{-2-\gamma} < S_2(i,i',k,k',m,m') < (p^\gamma \log p)^{-1}\} < C_2 \log p$, $\max_{1 \le i,i' \le p} \#\{(k,k',m,m') : S_2(i,i',k,k',m,m') > (p^\gamma \log p)^{-1}\} < C_2$, where $S_2(i,i',k,k',m,m') = T^{-4}\sum_{t=1}^T \sum_{t'=1}^T \sum_{s=1}^T \sum_{s'=1}^T |\operatorname{Cov}(u_{it}u_{kt}u_{it'}u_{k't'}, u_{i',s}u_{ms}u_{i',s'}u_{m's'})|$.
(iv) Tail behavior: There exist $r_2, r_3 > 1$ with $r_1^{-1} + r_2^{-1} + r_3^{-1} > 1$ and $b_1, b_2 > 0$ such that for each $i = 1, \ldots, p$, $k = 1, \ldots, K$ and any $s > 0$, $\mathbb{P}(|u_{it}| > s) \le \exp\{-(s/b_1)^{r_2}\}$ and $\mathbb{P}(|f_{tk}| > s) \le \exp\{-(s/b_2)^{r_3}\}$.

Condition 2.3 extends the standard assumptions for the factor analysis of large scale panel data or high-dimensional time series (Bai [8], Stock and Watson [86], Fan, Liao, and Wang [48]). Compared to existing conditions in the literature, we only require $\{\boldsymbol{u}_t, \mathbf{f}_t\}_{t \le T}$ to be weakly stationary rather than strictly stationary in (i) by carefully exploiting Davydov's inequality (Athreya and Lahiri [7]). Furthermore, in contrast to the traditional conditions for PCA (Jolliffe [65]) and factor models (Anderson and Rubin [6], Lawley and Maxwell [70], Anderson [4]), where either $u_{1t}, \ldots, u_{pt}$ are assumed to be independent with each other at each t or no temporal dependence across t is imposed on $\boldsymbol{u}_t$, (ii) and

(iii) together allow the error process to have both cross-sectional and temporal dependence. In fact, (iii) suggests that though the common factors explain most dependence within $\boldsymbol{y}_t$, the errors also account for some weak cross-sectional dependence. Assumptions similar to (iii) have been widely employed in the literature of high-dimensional statistics (Cai, Liu, and Xia [31], Fan, Liao, and Wang [48], Fan et al. [46]). While independent $u_{1t}, \ldots, u_{pt}$ easily satisfy (iii) for $C_2 = 2$, the inequalities in (iii) also hold under some weak dependence among $u_{1t}, \ldots, u_{pt}$. For example, if $u_{2t}, \ldots, u_{pt}$ are independent with each other and $u_{1t} = b_1 u_{2t} + b_2 u_{3t}$ for some non-zero constants b_1 and b_2, conditions in (iii) hold for $C_2 = 3$; also, if $S_1(i, i', k, m) < p^{-2-\gamma}$ and $S_2(i, i', k, k', m, m') < p^{-2-\gamma}$ for each i, i', k, k', m, m', (iii) holds for $C_2 = 2$ as well. Additionally, it is easy to see $||\boldsymbol{\Sigma}_u||_2 = O(1)$ from (iii), and together with Condition 2.2 they are the well-known *pervasiveness assumption*.

It is interesting to notice that the well-known Condition PC1 from Bai and Ng [13] restricts $\mathbf{F}$ to a subspace $\{\mathbf{F} \in \mathbb{R}^{T\times K} : T^{-1}\mathbf{F}^\top\mathbf{F} = \mathbf{I}_K\}$. However, for an arbitrary K-dimensional process under Condition 2.1, $T^{-1}\mathbf{F}^\top\mathbf{F}$ does not necessarily degenerate to its expected value $\mathbb{E}(T^{-1}\mathbf{F}^\top\mathbf{F}) = \mathrm{Var}(\mathbf{f}_1) = \mathbf{I}_K$. To satisfy this subspace restriction, one needs to rescale each realization of $\mathbf{F}$. Since the rescaling operator depends on the realization of $\mathbf{F}$, the rescaled processes no longer follow the original model of $\mathbf{f}_t$ if we assume any. This brings extra challenges to many applications. For example, in Section 4.2, this subspace restriction will prevent directly modeling $\mathbf{f}_t$ in (1.1) with some parametric models to forecast high-dimensional time series. In fact, we notice that this subspace restriction is stringent and can be replaced by the exponential tail bound on the difference between $T^{-1}\mathbf{F}^\top\mathbf{F}$ and its expectation $\mathbf{I}_K$. From the aforementioned well-known conditions, this bound can be easily established with the help of the τ-mixing coefficient as defined below.

Definition 2.1 (τ-mixing coefficient (Merlevède, Peligrad, and Rio [72])). For any real random variable X and σ-algebra $\mathcal{M}$, denote $\mathbb{P}_X$ the distribution of X and $\mathbb{P}_{X|\mathcal{M}}$ the conditional distribution of X on $\mathcal{M}$. The τ-mixing coefficient is defined by

$$\tau(\mathcal{M}, X) = \sup_{g\in\mathcal{L}_1(\mathbb{R})} \left| \int g(x)\mathbb{P}_{X|\mathcal{M}}(x) - \int g(x)\mathbb{P}_X(x) \right|,$$

where $\mathcal{L}_1(\mathbb{R})$ is the set of 1-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$.

Then, the τ-mixing coefficient of $\{f_{tk}\}$ for each $k = 1, \ldots, K$ is

$$\tau(T) = \sup_{j\geq 1} \frac{1}{j} \sup_{s>0, T+s\leq t_1<\cdots<t_j} \tau\left(\sigma(f_{tk}, t \leq s), (f_{t_1k}, \ldots, f_{t_jk})\right)$$

where $\sigma(f_{tk}, t \leq s)$ is the σ-algebra generated from $\{f_{tk}, t \leq s\}$. Note that, by Condition 2.3 (iv), for each $k = 1, \ldots, K$ and $t = 1, \ldots, T$,

$$Q(x) = \sup_{k,t} \inf\{s > 0 : \mathbb{P}(|f_{tk}^2| > s) \leq x\} = b_2^2\{\log(1/x)\}^{2/r_3}.$$

Thus, for $r_4 \in (0, 1)$ and any $x \geq 1$,

$$\tau(x) \leq 2\int_0^{2\alpha(x)} Q(u)du \leq 4b_2^2 r_4 \left\{\frac{r_3(1-r_4)}{2}\right\}^{2/r_3} \exp\left\{\frac{2}{r_3(1-r_4)}\right\} \{2\alpha(x)\}^{r_4},$$

which implies that $\mathbf{f}_t$ is τ-mixing by Condition 2.3 (ii). Then, following Theorem 1 in Merlevède, Peligrad, and Rio [72], with probability at least $1 - T^{-1}$,

$$\|T^{-1}\mathbf{F}^\top\mathbf{F} - \mathbf{I}_k\|_{\mathbb{F}}^2 \lesssim \frac{\log T}{T},$$

which is the desired assumption in place of the subspace restriction on $\mathbf{F}$.

3. Main results

Now we are in position to discuss the main results on non-asymptotic properties of the spectral decomposition of large Gram-type matrices based on (1.1) or (1.2). Continue to let $\mathbf{Y} = \mathbf{A}\mathbf{F}^\top + \mathbf{U}$, and we denote $T^{-1/2}\widehat{\boldsymbol{f}}_k$ the eigenvector corresponding to the kth largest eigenvalue of the right Gram matrix $\mathbf{Y}^\top\mathbf{Y}$ for $k = 1, \ldots, K$. Then, the loading matrix $\mathbf{A}$ can be estimated by $\widehat{\mathbf{A}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}$, where $\widehat{\mathbf{F}} = (\widehat{\boldsymbol{f}}_1, \ldots, \widehat{\boldsymbol{f}}_K)$. First, we have the following exponential tail bounds on the deviations $\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2$ and $\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\max}$.

Theorem 3.1 (Exponential tail bounds on the deviation between $\widehat{\mathbf{F}}$ and F). *Under Conditions 2.1-2.3, the deviation between $\widehat{\mathbf{F}}$ and $\mathbf{F}$ satisfies*

(i) *with probability at least* $1 - e^{-s}$,

$$T^{-1}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2 \lesssim \left(\frac{1}{p} + \frac{1}{T}\right)s^4;$$

(ii) $T^{-1}\mathbb{E}(\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2) \lesssim p^{-1} + T^{-1}$ *and* $T^{-2}\operatorname{Var}(\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2) \lesssim p^{-2} + T^{-2}$; *and*
(iii) *with probability at least* $1 - e^{-s}$,

$$\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\max} \lesssim \left(\frac{1}{\sqrt{p}} + \frac{1}{T}\right)(\log T)^{2/r_3}s.$$

For the approximate factor model, it has been shown that the mean squared error (MSE) $T^{-1}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2$ converges to zero when p and T diverge, thus $\widehat{\mathbf{F}}$ converges to $\mathbf{F}$ in probability (Bai and Ng [13], Fan, Liao, and Wang [48]). In Theorem 3.1, not only have we provided the non-asymptotic characterization on the MSE of $\widehat{\mathbf{F}}$ in the sense that the result holds for finite T and p, but also the convergence of $\widehat{\mathbf{F}}$ to $\mathbf{F}$ is established under a weaker condition on $\mathbf{F}$ compared to Condition PC1 in Bai and Ng [13] as discussed in Section 2.2. Theorem 3.1 reveals that the deviation between $\widehat{\mathbf{F}}$ and $\mathbf{F}$ is due to 1) the deviation between $\mathbf{F}$ and its projection onto subspace $\{\mathbf{F} \in \mathbb{R}^{T\times K} : T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K\}$, which is of rate $p^{-1} + T^{-2}$; and 2) the error for estimating this projection, which is of rate $p^{-2} + T^{-1}$. They lead to the non-asymptotic bound on $T^{-1}\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2$ in (i). In addition, $(p+T)^{-1}p\|\widehat{\mathbf{F}} - \mathbf{F}\|_{\mathbb{F}}^2$ enjoys a sub-exponential tail with the finite first and second moments from (ii).

Recall that both $\widehat{\mathbf{F}}$ and $\mathbf{F}$ have finite K columns. A by-product of Theorem 3.1 is an exponential tail bound on the deviation between the $T^{-1/2}$-scaled kth columns of $\widehat{\mathbf{F}}$, *i.e.*, the kth eigenvector of the right Gram matrix, and its counterpart in $\mathbf{F}$. That is, with probability at least $1 - e^{-s}$, for each $k = 1, \ldots, K$,

$$T^{-1}\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2 \lesssim \left(\frac{1}{p} + \frac{1}{T}\right)s^4.$$

Therefore, $(p+T)^{-1}p\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2$ also admits a sub-exponential tail with the finite first and second moments, which are similar to (ii) in Theorem 3.1.

Using the max norm, the error rate remains the same for recovering the projection since it is of finite dimension. On the other hand, the ℓ_∞-deviation between $\mathbf{F}$ and its projection is of rate $(p^{-1/2} + T^{-1})(\log T)^{2/r_3}$, where $\log T$ is due to the maximum inequality to control the maximum among TK entries in $\mathbf{F}$. Result in (iii) provides a non-asymptotic entry-wise bound on the deviation between $\widehat{\mathbf{F}}$ and

F. For each $t = 1, \ldots, T$ and $k = 1, \ldots, K$, $|\widehat{f}_{tk} - f_{tk}|(p^{-1/2} + T^{-1})^{-1}(\log T)^{-2/r_3}$ displays a sub-exponential tail. Thus, following the similar argument in (ii), $(p^{-1/2} + T^{-1})^{-1}(\log T)^{-2/r_3}|\widehat{f}_{tk} - f_{tk}|$ also has the finite first and second moments for all p and T. By Condition 2.3, f_{tk} has the finite first and second moments and so does $\widehat{f}_{tk}$ whenever $(p^{-1/2} + T^{-1})(\log T)^{2/r_3} = O(1)$ due to the triangle inequality. This nontrivial result makes it possible to further model the K-dimensional latent process parametrically; see Section 4.2 for more details.

Next, in Theorem 3.2, we establish the Berry-Esseen type bound for each of the K eigenvectors of the right Gram matrix. It provides the approximation error rate to the distribution of the standardized deviation between $\widehat{\boldsymbol{f}}_k$ and $\boldsymbol{f}_k$ by the standard normal distribution for each k.

Theorem 3.2 (Berry-Esseen Type Bound for $||\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k||_2^2$). *Under Conditions* 2.1-2.3, *for each* $k = 1, \ldots, K$, *we have*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left\{ \frac{\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2 - \mathbb{E}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2)}{\mathrm{Var}^{1/2}(\|\widehat{\boldsymbol{f}}_k - \boldsymbol{f}_k\|_2^2)} \le x \right\} - \Phi(x) \right| \lesssim \frac{\log T}{\sqrt{p}} + \frac{1}{\sqrt{T}},$$

where $\Phi(x)$ *is the cumulative distribution function of the standard normal distribution.*

Theorem 3.2 sheds lights on drawing inference on the leading eigenvectors of the covariance matrix for non *i.i.d.* data, which is detailed in Section 4.3. For *i.i.d.* data, the traditional rate in the Berry-Esseen bound for Gaussian approximation is $T^{-1/2}$ (Chan and Wierman [35], Callaert and Janssen [33]). In Theorem 3.2, $p^{-1/2}$ and $T^{-1/2}$ are due to the uncertainty from $\mathbf{f}_t$ and $\boldsymbol{u}_t$ for computing $\widehat{\boldsymbol{f}}_k$. In addition, the dependence in $\mathbf{f}_t$ leads to the extra $\log T$ in the bound, which has been observed in the literature (Hörmann [55], Jirak [61]).

In the rest of this section, we will study non-asymptotic properties of the eigenvalues of $\mathbf{Y}^\top\mathbf{Y}$. Although the spectral structure of the expected right Gram matrix $\mathbb{E}(\mathbf{Y}^\top\mathbf{Y})$ differs from that of the expected left Gram matrix $\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$, it is interesting to notice that $\mathbf{Y}^\top\mathbf{Y}$ and $\mathbf{Y}\mathbf{Y}^\top$ share the common non-zero eigenvalues. Hence, we first consider $\mathbf{Y}\mathbf{Y}^\top$, which is conveniently the sample covariance matrix scaled by T. Denote $\{\lambda_i\}_{i=1}^p$ and $\{\boldsymbol{w}_i\}_{i=1}^p$ the eigenvalues (in decreasing order) and corresponding eigenvectors of $\boldsymbol{\Sigma} = T^{-1}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)$, and let $\{\widehat{\lambda}_i\}_{i=1}^p$ and $\{\widehat{\boldsymbol{w}}_i\}_{i=1}^p$ be the eigenvalues (in decreasing order) and corresponding eigenvectors of $\widehat{\boldsymbol{\Sigma}} = T^{-1}\mathbf{Y}\mathbf{Y}^\top$. We establish the non-asymptotic characterization of $\widehat{\lambda}_i$ relative to λ_i as follows.

Theorem 3.3 (Non-asymptotic characterization of $\widehat{\lambda}_i$'s relative to λ_i's). *Under Conditions* 2.1-2.3, *there exist positive constants* C *and* c *that only depend on* $\boldsymbol{u}_t$ *such that the following results hold.*

(i) *If* $p < T$, *with probability at least* $1 - e^{-s}$,

$$|\widehat{\lambda}_i/\lambda_i - 1| \le \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}, \qquad i = 1, \ldots, K,$$

$$|\widehat{\lambda}_i/\lambda_i - 1| \le \frac{C\sqrt{p}}{\sqrt{T}} + \frac{c}{\sqrt{T}}\sqrt{s}, \qquad i = K+1, \ldots, p.$$

(ii) *If* $p \ge T$, *with probability at least* $1 - e^{-s}$,

$$|\widehat{\lambda}_i/\lambda_i - 1| \le \frac{C}{\sqrt{T}} + \frac{c}{\sqrt{pT}}\sqrt{s}, \qquad i = 1, \ldots, K,$$

$$\widehat{\lambda}_i/\lambda_i \ge \frac{p}{T} - C\sqrt{\frac{p}{T}} - \frac{c}{\sqrt{T}}\sqrt{s}, \qquad i = K+1, \ldots, T,$$

$$\widehat{\lambda}_i/\lambda_i \le \frac{p}{T} + C\sqrt{\frac{p}{T}} + \frac{c}{\sqrt{T}}\sqrt{s}, \qquad i = K+1, \ldots, T.$$

Taking $s = \log T$, Theorem 3.3 implies that the first K eigenvalues of the scaled left Gram matrices, *i.e.*, the sample covariance matrix, $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K$ converges to the corresponding eigenvalues of $\mathbf{\Sigma}$ in probability. When $p < T$, the relative errors of the remaining $p - K$ eigenvalues to their population counterparts are bounded by $T^{-1/2}p^{1/2}$ in probability. By Condition 2.1, λ_i is bounded for $i > K$. Thus, the bound on relative error $|\widehat{\lambda}_i/\lambda_i - 1|$ is the same as that of deviation $|\widehat{\lambda}_i - \lambda_i|$ for $i > K$. That is, eigenvalues of $\widehat{\mathbf{\Sigma}}$ converge to those of $\mathbf{\Sigma}$ only if $p/T \to 0$. This agrees with the well known convergence of $\widehat{\mathbf{\Sigma}}$ to $\mathbf{\Sigma}$ in low dimension for *i.i.d.* data (Bunea and Xiao [29], Bien, Bunea, and Xiao [24]).

Different lessons are learned when $p > T$. As $\widehat{\mathbf{\Sigma}}$ is not of full-rank, $\widehat{\lambda}_i$'s with $i > K$ consist of at most $T - K$ non-zeros and at least $p - T$ zeros. For a legitimate covariance $\mathbf{\Sigma}$, at least $p - T$ eigenvalues of $\widehat{\mathbf{\Sigma}}$ are biased for estimating their population counterparts. In addition, the non-zero eigenvalues of $\widehat{\mathbf{\Sigma}}$ could also be biased. For *i.i.d.* data with unit variance and p proportional to T, it is known that non-zero eigenvalues of the sample covariance matrix are spread out and bounded by $(1 - p^{1/2}T^{-1/2})^2$ and $(1 + p^{1/2}T^{-1/2})^2$ (Stein [85], James and Stein [59], Bai and Yin [16], Johnstone and Paul [64]), which explains the bias in non-zero eigenvalues of the sample covariance matrix compared to their population counterparts (Bai and Yin [16], Baik, Arous, and Péché [19], Johnstone and Paul [64]). In contrast, the low-rank structure in factor models provides better understanding on eigenvalues of $\widehat{\mathbf{\Sigma}}$. Consider a factor model with $\boldsymbol{u}_t$ assumed to be white noise, Lam and Yao [69] focused on the cross covariance matrix $\mathbf{M} = \sum_{h=1}^{h_0} \mathbf{\Sigma}(h)\mathbf{\Sigma}(h)^\top$, where $\mathbf{\Sigma}(h)$ is the autocovariance matrix of $\boldsymbol{y}_t$ at lag h. They remarked that asymptotically, spiked eigenvalues of the sample cross covariance matrix converge to the corresponding population eigenvalues, while the non-spiked eigenvalues, although may not converge, are bounded by the ratio of p and T. Theorem 3.3 (ii) provides a non-asymptotic characterization of their remarks. First, we confirm that, as expected, $\widehat{\lambda}_i$ fails to converge for $i > K$ if p/T diverges. Also, the non-asymptotic bound in Theorem 3.3 shows that the ratio between $\mathbb{E}(\widehat{\lambda}_i)$ and λ_i is bounded above by $2\sqrt{\pi cp/T}\Phi(2^{-1/2}c^{-1}C\sqrt{p})$ for any given p and T. Furthermore, the non-asymptotic bound of $\widehat{\lambda}_i/\lambda_i$ provide a characterization on the closeness between $\mathbb{E}(\widehat{\lambda}_i/\lambda_i)$ and 1 for $i = 1, \ldots, K$. It is easy to see from Theorem 3.3 that the deviation between $\mathbb{E}(\widehat{\lambda}_i/\lambda_i)$ and 1 is bounded above by $2\sqrt{\pi c/(pT)}\Phi(2^{-1/2}c^{-1}C\sqrt{p})$. This reflects the asymptotic unbiasedness of $\widehat{\lambda}_i$ for $i = 1, \ldots, K$.

Our focus on the non-asymptotic behavior of $\widehat{\lambda}_i/\lambda_i$ in Theorem 3.3 is fueled in part by the efforts on establishing the convergence rate of estimated number of latent factors in PCA using the consecutive eigenvalues of sample covariance matrix, such as the eigenvalue-ratio test (Lam and Yao [69], Ahn and Horenstein [2], Fan, Liao, and Wang [48]), which is detailed in Section 4.1. Parallel to the above non-asymptotic results, in the literature, the consistency and asymptotic normality of $\widehat{\lambda}_i/\lambda_i$ have been documented. For example, in Wang and Fan [88], the authors considered a noiseless factor model with arbitrary factor strengths and sub-Gaussian factors, which allows the spiked eigenvalue to be with any rate in p. They showed that eigenvalues of $\widehat{\mathbf{\Sigma}}$ are asymptotically unbiased if $pT^{-1}\lambda_i^{-1}$ converge to zero for $i = 1, \ldots, K$. Also, the authors established the asymptotic normality of $\widehat{\lambda}_i/\lambda_i - 1$ upon removing the bias. Lately, Cai, Han, and Pan [30] considered a p-dimensional spiked covariance model with K spiked eigenvalues that are separated from others, where K is potentially divergent. Specifically, they focused on data generated from $(p + K) \times T$ *i.i.d* random variables with zero mean, unit variance, and finite fourth moment, loaded on a $p \times (p + K)$ deterministic matrix. The authors employed the Stieltjes transform method to study the contribution of non-spiked eigenvalues to the spiked ones, and therefore obtained a refined characterization of bias of $\widehat{\lambda}_i/\lambda_i - 1$ as well as the asymptotic normality of $\widehat{\lambda}_i/\lambda_i - 1$. In terms of modeling, if $\boldsymbol{u}_t$ is further modeled by $\mathbf{C}\boldsymbol{f}_t$ for some $\mathbf{C}$ orthogonal to $\mathbf{A}$ and

$\boldsymbol{f}_t$ is sub-Gaussian, (1.2) reduces to a special case that coincides with the model in Wang and Fan [88], where the spiked eigenvalues are all at the same rate of p. Recall that $\lambda_i = O(p)$ under Condition 2.2, so that $pT^{-1}\lambda_i^{-1}$ always converges to zero for (1.2). Thus, Theorem 3.3 gives a similar result on the asymptotic unbiasedness of $\widehat{\lambda}_i$ as Wang and Fan [88]. In addition, Theorem 3.3 (ii) provides a finite sample view of $\widehat{\lambda}_i/\lambda_i$ by showing its non-asymptotic sub-Gaussian tail for $i = 1, \ldots, K$. On the other hand, if f_{tk} and u_{it} are *i.i.d* for each $t = 1, \ldots, K$, $i = 1, \ldots, p$, and $k = 1, \ldots, K$ in (1.1) while the $p \times (p+K)$ deterministic matrix in Cai, Han, and Pan [30] has full rank, our model agrees with the one in Cai, Han, and Pan [30]. Hence, it is possible to establish the non-asympotitic results in Theorem 3.3 by exploring the techniques employed to establish the consistency of $\widehat{\lambda}_i/\lambda_i$ in Cai, Han, and Pan [30], which we will leave to future efforts.

4. Applications in high-dimensional statistics

To demonstrate results in Section 3, we consider a number of interesting and widely studied applications in high-dimensional statistics, including the estimation of the number of latent factors in factor models and related machine learning problems, the estimation and forecasting of high-dimensional time series, the spectral properties of large sample covariance matrix such as perturbation bounds and inference on the spectral projectors, and the low-rank matrix denoising from dependent data.

4.1. Estimation of the number of latent factors

In high-dimensional factor models or machine learning problems such as PCA, it is necessary to choose the number of latent factors or principal components K before recovering the loading matrix and factors or computing the principal components and scores. Traditional methods to estimate K include, for example, the likelihood ratio test and the scree plot (Jolliffe [65]). For the high-dimensional data with large covariance matrix, eigenvalues of the sample covariance matrix or their variants have been utilized and the estimation is consistent under certain separation conditions of the first K eigenvalues from the remains. A popular approach is based on the ratio of consecutive eigenvalues (Lam and Yao [69], Ahn and Horenstein [2], Fan, Liao, and Wang [48]),

$$\widehat{K} = \operatorname{argmax}_{1 \le i < \min(p,T)} \frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}} \tag{4.1}$$

where $\widehat{\lambda}_i$ is the ith eigenvalue of $T^{-1}\mathbf{Y}\mathbf{Y}^\top$; while, other methods are based on the eigenvalue differences (Onatski [76]) or the cumulative magnitude of eigenvalues (Bai and Ng [11]).

Under the pervasiveness assumption, *i.e.*, Condition 2.2 and (iii) in Condition 2.3, the consistency of $\widehat{K}$ has been established (Lam and Yao [69], Fan, Liao, and Wang [48]). However, the rate of the probability of consistent estimation has not been fully explored. Theorem 3.3 sheds light on characterizing this rate. In fact, from Theorem 3.3, $\widehat{\lambda}_K/\widehat{\lambda}_{K+1}$ is of the order $O_p(p)$ when $p < T$ and $O_p(T)$ when $p > T$. In contrast, $\widehat{\lambda}_i/\widehat{\lambda}_{i+1}$ is $O_p(1)$ for $i \neq K$. As an application, Theorem 4.1 establishes the non-asymptotic lower bound of the probability of estimating the correct number of factors.

Theorem 4.1. *Under Conditions* 2.1-2.3, *given* $\mathbf{Y}$ *from* (1.1) *or* (1.2), $\widehat{K}$ *defined in* (4.1) *satisfies*

$$\mathbb{P}(\widehat{K} = K) \ge 1 - \exp(-\{C_1\sqrt{\max(p,T)} - C_2\sqrt{\min(p,T)}\}^2), \tag{4.2}$$

where

$$C_1 = \frac{1}{c}\left[1 - \left\{\frac{\max(p,T)\lambda_{K+1}}{T\lambda_K} \max_{1\le i<\min(p,T),i\neq K} \frac{\lambda_i}{\lambda_{i+1}}\right\}^{1/4}\right],$$

and $C_2 = c^{-1}C$, *with* C *and* c *defined in Theorem* 3.3.

As mentioned in Theorem 3.3, C and c are positive constants that only depend on $\boldsymbol{u}_t$ so that $C_2 > 0$ is independent of p and T. Under Conditions 2.1 and 2.2, $\lambda_i = O(p)$ for $i = 1, \ldots, K$ and $\lambda_i = O(1)$ for $i > K$ so that $C_1 > 0$ for sufficiently large p and T. On the right hand side of (4.2), $C_2\sqrt{\min(p,T)}$ is smaller than $C_1\sqrt{\max(p,T)}$ whenever $p \ll T$ or $p \gg T$, so the lower bound is governed by $C_1\sqrt{\max(p,T)}$. It is easy to see that C_1 is large if both λ_{K+1}/λ_K and $\max_{1\le i<\min(p,T),i\neq K}\lambda_i/\lambda_{i+1}$ are small. That is, it is easy to estimate K if the spiked eigenvalues $\lambda_1, \ldots, \lambda_K$ are close to each other and so do the non-spiked eigenvalues $\lambda_{K+1}, \ldots, \lambda_p$. Otherwise, if λ_i/λ_{i+1} is large for some $i \neq K$, C_1 will be small so that the lower bound on the right hand side of (4.2) will be away from 1 and implies a more challenging K to be estimated.

When p and T are close, $C_2\sqrt{\min(p,T)}$ is not negligible. Notice that the lower bound in (4.2) can be written as $1 - \exp\{-C_1^2(1 - C'p^{1/2}T^{-1/2})^2T\}$ for some positive constant C' given $p < T$. When $C_1^2(1 - C'p^{1/2}T^{-1/2})^2$ is small, a large T is preferable to drive the lower bound close to 1. If $p \ge T$, the lower bound in (4.2) can be written as $1 - \exp\{-C_1^2(1 - C'T^{1/2}p^{-1/2})^2p\}$ and similarly, a large p is preferable to make the lower bound approaching 1.

An alternative to $\widehat{K}$, proposed by Onatski [76], is to use the difference of consecutive eigenvalues. That is, for given $\delta > 0$ and predetermined L, one defines

$$\widehat{K}_d = \max\{i \le L : \widehat{\lambda}_i - \widehat{\lambda}_{i+1} \ge \delta\}. \tag{4.3}$$

Similar to Theorem 4.1, we have the following result.

Theorem 4.2. *Under Conditions* 2.1-2.3, *given* $\mathbf{Y}$ *from* (1.1) *or* (1.2), $\widehat{K}_d$ *in* (4.3) *satisfies*

$$\mathbb{P}(\widehat{K}_d = K) \ge 1 - \sum_{i=1}^{K+1} \exp(-\{C_{1i}\sqrt{T} - C_2\sqrt{p}\}^2),$$

where $C_{1i} = (2c)^{-1}(\lambda_i - \lambda_{i+1} - \delta)$ *for* $i = 1, \ldots, K$, $C_{1,K+1} = c^{-1}(\delta - \lambda_{K+2} + \lambda_{K+1})$, *and* $C_2 = c^{-1}C$, *with* C *and* c *defined in Theorem* 3.3.

Under the pervasiveness assumption, Onatski [76] established the consistency of $\widehat{K}_d$ when p is proportional to T. Theorem 4.2 relaxes the restriction on p and T and provides the non-asymptotic characterization of the probability of consistent estimation of K by $\widehat{K}_d$. It suggests that, for carefully selected δ such that $\delta > \lambda_{K+2} - \lambda_{K+1}$, $\widehat{K}_d$ and $\widehat{K}$ have similar rates of the probability of consistent estimation. However, $\widehat{K}_d$ is not tuning free compared to $\widehat{K}$. Onatski [76] proposed a data-driven procedure to determine δ. Specifically, an iterative procedure was employed to alternatively update δ and $\widehat{K}_d$ until convergence. Note that $\lambda_{K+2} = \lambda_{K+1}$ if $u_{1t}, \ldots, u_{pt}$ are identical. In this case, an appropriate δ can be easily found. Otherwise, more numerical iterations are required. Sometimes, $\widehat{K}_d$ may perform better than $\widehat{K}$ in practice, which can be explained using the non-asymptotic results from Theorems 4.1 and 4.2. Consider a special case where $p > T$, $K = 1$, $\lambda_1 = p$, and $\lambda_2 = \cdots = \lambda_p = 1$. The lower bound for $\widehat{K}$ in Theorem 4.1 is $1 - \exp(-\{c^{-1}(1 - T^{-1/4})\sqrt{p} - c^{-1}C\sqrt{T}\}^2)$ while the lower bound for $\widehat{K}_d$ in

Theorem 4.2 is $1-\exp(-\{(2c)^{-1}(p-1-\delta)\sqrt{T}-c^{-1}C\sqrt{p}\}^2)-\exp(-\{(2c)^{-1}\delta\sqrt{T}-c^{-1}C\sqrt{p}\}^2)$. For a divergent p and constant T, $\widehat{K}_d$ outperforms $\widehat{K}$ in terms of a higher rate of the probability of consistent estimation whenever $C>1-T^{-1/4}$.

Different from the consecutive eigenvalue based approaches, the information criterion has also been used to estimate K. Some of them can be interpreted as a penalized cumulative magnitude of eigenvalues, such as

$$\mathbb{PC}(k)=\left\{\frac{1}{pT}\sum_{i\geq k}\widehat{\lambda}_i+k\widehat{\sigma}^2\frac{p+T}{pT}\log\left(\frac{pT}{p+T}\right)\right\},$$

where $\widehat{\sigma}^2$ is some consistent estimate of $(pT)^{-1}\sum_{i=1,t=1}^{p,T}\mathbb{E}(u_{it}^2)$ (Bai and Ng [11]). Then, K is estimated by $\widehat{K}_m=\operatorname{argmin}_{k\leq L}\mathbb{PC}(k)$ for some predetermined L. Bai and Ng [11] suggested that $\widehat{\sigma}^2$ can be replaced by $(pT)^{-1}\sum_{i>L}\widehat{\lambda}_i$ in practice and the penalty $(pT)^{-1}(p+T)\log((p+T)^{-1}pT)$ can be replaced by $(pT)^{-1}(p+T)\log(\min(p,T))$ or $\min(p,T)^{-1}\log(\min(p,T))$. They also showed the consistency of $\widehat{K}_m$ when $\widehat{\sigma}^2$ is consistent and the penalty shrinks to zero as p and T diverge. Notice that $\widehat{K}_m$ is entirely based on the empirical distribution of $\widehat{\lambda}_i$ for $i=1,\ldots,p$. Thus, its non-asymptotic properties such as the rate of the probability of consistent estimation may also be established using Theorem 3.3, which we leave to the future work.

4.2. Estimation and forecasting of high-dimensional time series

Making forecasts based on high-dimensional time series arises frequently in econometrics, financial analysis, and meteorology. Suppose we observe $\mathbf{Y}\in\mathbb{R}^{p\times T}$, where each entry y_{it} follows (1.1) and the zero mean K-dimensional latent process $\mathbf{f}_t$ is governed by parametric models satisfying Conditions 2.1 and 2.3. For example, Chen, Wang, and Wu [36] considered a model similar to (1.1) with $\mathbf{f}_t$ following an autoregressive model whose parameters are estimated for predicting $y_{i,s}$ with $s>T$.

As an application of Theorem 3.1, we show the consistency on estimating the moments of $\mathbf{f}_t$ using the spectral decomposition of $\mathbf{Y}^\top\mathbf{Y}$, which guarantees the consistency of moment-based estimators to parameters of a large realm of parametric models for $\mathbf{f}_t$. Let the sample autocovariance function (Brockwell, Davis, and Fienberg [27]) of $\mathbf{f}_t$ be

$$\widehat{\mathbf{\Gamma}}(h,\mathbf{f}_t)=\frac{1}{T}\sum_{t=1}^{T-|h|}(\mathbf{f}_{t+|h|}-\bar{\mathbf{f}})(\mathbf{f}_t-\bar{\mathbf{f}})^\top,$$

where $\bar{\mathbf{f}}=T^{-1}\sum_{t=1}^{T}\mathbf{f}_t$. Also, denote the sample autocovariance function of $\widehat{\mathbf{f}}_t$, the tth row of $\widehat{\mathbf{F}}$, by

$$\widehat{\mathbf{\Gamma}}(h,\widehat{\mathbf{f}}_t)=\frac{1}{T}\sum_{t=1}^{T-|h|}(\widehat{\mathbf{f}}_{t+|h|}-\bar{\widehat{\mathbf{f}}})(\widehat{\mathbf{f}}_t-\bar{\widehat{\mathbf{f}}})^\top,$$

where $\bar{\widehat{\mathbf{f}}}=T^{-1}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t$. In Theorem 4.3, we show that the sample autocovariance function of $\mathbf{f}_t$ can be consistently recovered by that of $\widehat{\mathbf{f}}_t$.

Theorem 4.3. *Under Conditions 2.1-2.3, given* $\mathbf{Y}$ *from* (1.1) *or* (1.2), $\widehat{\mathbf{\Gamma}}(h,\mathbf{f}_t)$ *and* $\widehat{\mathbf{\Gamma}}(h,\widehat{\mathbf{f}}_t)$ *defined above satisfy, with probability at least* $1-e^{-s}$,

$$\|\widehat{\mathbf{\Gamma}}(h,\widehat{\mathbf{f}}_t)-\widehat{\mathbf{\Gamma}}(h,\mathbf{f}_t)\|_{\mathbb{F}}^2\lesssim\frac{1}{T}\left(\frac{1}{p}+\frac{1}{T}\right)s$$

for each $h = -T+1, \ldots, 0, \ldots, T-1$.

Notice that both $\mathbf{F}$ and $\widehat{\mathbf{F}}$ are $T \times K$ matrices. As a direct corollary of Theorem 4.3, we can establish the concentration inequality for recovering the temporal dependence structure on each dimension of $\mathbf{f}_t$. For each $k = 1, \ldots, K$, denote the sample autocovariance function of $\{f_{tk} : t \geq 1\}$ as

$$\widehat{\gamma}(h, f_{tk}) = T^{-1} \sum_{t=1}^{T-|h|} (f_{t+|h|,k} - \bar{f}_k)(f_{tk} - \bar{f}_k)^\top,$$

where $\bar{f}_k = T^{-1} \sum_{t=1}^{T} f_{tk}$, and let the sample autocovariance function of $\{\widehat{f}_{tk} : t \geq 1\}$ be

$$\widehat{\gamma}(h, \widehat{f}_{tk}) = T^{-1} \sum_{t=1}^{T-|h|} (\widehat{f}_{t+|h|,k} - \bar{\widehat{f}}_k)(\widehat{f}_{tk} - \bar{\widehat{f}}_k)^\top,$$

where $\bar{\widehat{f}}_k = T^{-1} \sum_{t=1}^{T} \widehat{f}_{tk}$. From Theorem 4.3, with probability at least $1 - e^{-s}$, we have

$$|\widehat{\gamma}(h, \widehat{f}_{tk}) - \widehat{\gamma}(h, f_{tk})|^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right) s$$

for each $h = -T+1, \ldots, 0, \ldots, T-1$. Similarly, denote the sample autocorrelation function (ACF; Brockwell, Davis, and Fienberg [27]) of $\{f_{tk} : t \geq 1\}$ by $\widehat{\rho}(h, f_{tk}) = \{\widehat{\gamma}(0, f_{tk})\}^{-1} \widehat{\gamma}(h, f_{tk})$ and the sample partial autocorrelation function (PACF) by $\widehat{\Psi}(0, f_{tk}) = 1$ and $\widehat{\Psi}(h, f_{tk})$ being the hth entry of $\widehat{\boldsymbol{\Psi}}(f_{tk})$ where $\widehat{\boldsymbol{\Psi}}(f_{tk}) = \widehat{\mathbf{R}}_h^{-1}(f_{tk}) \widehat{\boldsymbol{\rho}}_h(f_{tk})$ with $\widehat{\mathbf{R}}_h(f_{tk}) = \{\widehat{\rho}((i-j), f_{tk})\}_{i,j=1}^{h}$ and $\widehat{\boldsymbol{\rho}}_h(f_{tk}) = (\widehat{\rho}(1, f_{tk}), \ldots, \widehat{\rho}(h, f_{tk}))^\top$. Likewise, we denote the sample ACF of $\{\widehat{f}_{tk} : t \geq 1\}$ by $\widehat{\rho}(h, \widehat{f}_{tk}) = \{\widehat{\gamma}(0, \widehat{f}_{tk})\}^{-1} \widehat{\gamma}(h, \widehat{f}_{tk})$, let the sample PACF of $\{\widehat{f}_{tk} : t \geq 1\}$ be $\widehat{\Psi}(0, \widehat{f}_{tk}) = 1$, and let $\widehat{\Psi}(h, \widehat{f}_{tk})$ be the hth entry of $\widehat{\boldsymbol{\Psi}}(\widehat{f}_{tk})$, where $\widehat{\boldsymbol{\Psi}}(\widehat{f}_{tk}) = \widehat{\mathbf{R}}_h^{-1}(\widehat{f}_{tk}) \widehat{\boldsymbol{\rho}}_h(\widehat{f}_{tk})$, $\widehat{\mathbf{R}}_h^{-1}(\widehat{f}_{tk}) = \{\widehat{\rho}((i-j), \widehat{f}_{tk})\}_{i,j=1}^{h}$ and $\widehat{\boldsymbol{\rho}}_h(\widehat{f}_{tk}) = (\widehat{\rho}(1, f_{tk}), \ldots, \widehat{\rho}(h, \widehat{f}_{tk}))^\top$. From Theorem 4.3, we have the following results.

Theorem 4.4. *Under Conditions* 2.1-2.3, *given* $\mathbf{Y}$ *from* (1.1) *or* (1.2), *for each* $k = 1, \ldots, K$ *and* $h = -T+1, \ldots, T-1$, *with probability at least* $1 - e^{-s}$,

$$|\widehat{\rho}(h, \widehat{f}_{tk}) - \widehat{\rho}(h, f_{tk})|^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right) s,$$

$$|\widehat{\Psi}(h, \widehat{f}_{tk}) - \widehat{\Psi}(h, f_{tk})|^2 \lesssim \frac{1}{T}\left(\frac{1}{p} + \frac{1}{T}\right) s.$$

Theorem 4.4 shows that sample ACF and PACF of $\{f_{tk} : t \geq 1\}$ can be consistently recovered by those of $\{\widehat{f}_{tk} : t \geq 1\}$. In addition, Theorem 4.4 implies that the sample ACF of f_{tk} and $\widehat{f}_{tk}$ have the common asymptotic distribution. Similar conclusions are also true for the sample PACF. These results will have wide applications in modeling and forecasting high-dimensional time series by (1.1) along a broad class of parametric models on $\mathbf{f}_t$. For instance, for the autoregression models, the sample PACF's give the Yule-Walker estimator to the autoregressive coefficients; and for the moving average models, the innovation estimator, which is computed from the sample ACF's, can be employed to estimate the moving average coefficients.

4.3. Spectral properties of large sample covariance matrices

Extending results in Section 3 on eigenvectors $\widehat{\boldsymbol{f}}_k$ of the scaled right Gram matrix $T^{-1}\mathbf{Y}^\top\mathbf{Y}$, we study eigenvectors of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{w}}_i$ for $i = 1, \ldots, p$. First, as an application of Theorem 3.1, we characterize the deviation between $\widehat{\boldsymbol{w}}_i$ and $\boldsymbol{w}_i$ in Theorem 4.5.

Theorem 4.5. *Under Conditions* 2.1-2.3, *given* $\mathbf{Y}$ *from* (1.1) *or* (1.2), $\mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2) \lesssim p^{-1} + T^{-1}$ *and* $\mathrm{Var}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2) \lesssim p^{-2} + T^{-2}$ *for each* $i = 1, \ldots, K$.

From Theorem 4.5, the first K eigenvectors of the sample covariance matrix converge to those of $\boldsymbol{\Sigma}$ in probability. Together with Theorems 3.3, we establish the consistency on estimating the spectral structure corresponding to the first K eigenvalues of $\boldsymbol{\Sigma}$ specified by (1.3). Notice that no restrictions on p and T are imposed on this consistency. Recall that $T\widehat{\boldsymbol{\Sigma}} = \mathbf{A}\mathbf{F}^\top\mathbf{F}\mathbf{A}^\top + \mathbf{A}\mathbf{F}^\top\mathbf{U} + \mathbf{U}^\top\mathbf{F}\mathbf{A}^\top + \mathbf{U}\mathbf{U}^\top$. By Lemmas C.1, C.9, and C.11 in the Supplementary Material (Zhang, Zhou, and Wang [93]), $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \le \|\mathbf{A}(T^{-1}\mathbf{F}^\top\mathbf{F} - \mathbf{I})\mathbf{A}^\top\|_2 + \|T^{-1}\mathbf{A}\mathbf{F}^\top\mathbf{U}\|_2 + \|T^{-1}\mathbf{U}^\top\mathbf{F}\mathbf{A}^\top\|_2 + \|T^{-1}\mathbf{U}\mathbf{U}^\top - \boldsymbol{\Sigma}_u\|_2 \lesssim T^{-1/2}p + T^{-1/2}p\sqrt{s}$ with probability at least $1 - e^{-s}$ for any $s > 0$. Thus, from the Davis-Kahan Theorem (Davis and Kahan [41], Yu, Wang, and Samworth [90], Fan, Wang, and Zhong [50], Zhang, Cai, and Wu [91]) and Condition 2.1, we have the following corollary.

Corollary 4.1. *Given* $\mathbf{Y}$ *from* (1.1) *or* (1.2), *let* $\Theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i) = \cos^{-1}(\widehat{\boldsymbol{w}}_i^\top\boldsymbol{w}_i)$ *be the angle between* $\widehat{\boldsymbol{w}}_i$ *and* $\boldsymbol{w}_i$. *Under Conditions* 2.1-2.3, *for each* $i = 1, \ldots, K$,

$$\mathbb{E}\{\sin\Theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i)\} \lesssim \frac{\mathbb{E}(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2)}{\min_{j\neq i}|\lambda_j - \lambda_i|} \lesssim T^{-1/2}.$$

Moreover, if $\widehat{\boldsymbol{w}}_i^\top\boldsymbol{w}_i \ge 0$, *then* $\mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2) \lesssim T^{-1/2}$ *for each* $i = 1, \ldots, K$.

Corollary 4.1 gives a similar result to the Davis-Kahan Theorem in low dimension. However, when $p > T$, as shown in Theorem 3.3, not all eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ necessarily converge to those of $\boldsymbol{\Sigma}$ and neither does $\widehat{\boldsymbol{\Sigma}}$ converge to $\boldsymbol{\Sigma}$. Then, the Davis-Kahan Theorem cannot be directly applied to $\widehat{\boldsymbol{\Sigma}}$. Instead, with the low-rank structure in (1.2), we can establish similar results for an alternative estimator to $\boldsymbol{\Sigma}$. We start with eigenvectors corresponding to the first K largest eigenvalues of $\mathbf{Y}^\top\mathbf{Y}$, *i.e.*, the PCA estimator to the latent factor matrix and loading matrix. If we further assume that $u_{1t}, \ldots, u_{pt}$ are uncorrelated for each t, $\boldsymbol{\Sigma}$ in (1.3) can be estimated by $\widehat{\boldsymbol{\Sigma}}_{\mathrm{PCA}} = \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top + \widehat{\boldsymbol{\Sigma}}_u$, where $\widehat{\mathbf{A}}$ is defined in Section 3, $\widehat{\boldsymbol{\Sigma}}_u$ is a diagonal matrix with diagonal entries $\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_p^2$, $\widehat{\sigma}_i^2 = T^{-1}\sum_{t=1}^T \widehat{u}_{it}^2$ for $i = 1, \ldots, p$ and $\widehat{u}_{it}$ is the entry in the ith row and tth column of $\widehat{\mathbf{U}} = \mathbf{Y} - \widehat{\mathbf{A}}\widehat{\mathbf{F}}^\top$. Then, similar to Corollary 4.1, we have the following result.

Corollary 4.2. *Given* $\mathbf{Y}$ *from* (1.1) *or* (1.2), *let* $\Theta(\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}, \boldsymbol{w}_i) = \cos^{-1}(\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}^\top\boldsymbol{w}_i)$ *be the angle between* $\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}$ *and* $\boldsymbol{w}_i$, *where* $\widetilde{\boldsymbol{w}}_{i,\mathrm{PCA}}$ *be the eigenvector corresponding to the* i*th largest eigenvalue of* $\widehat{\boldsymbol{\Sigma}}_{\mathrm{PCA}}$. *Then, under Conditions* 2.1-2.3, *for each* $i = 1, \ldots, K$,

$$\mathbb{E}\{\sin\Theta(\widehat{\boldsymbol{w}}_{i,\mathrm{PCA}}, \boldsymbol{w}_i)\} \lesssim \frac{\mathbb{E}(\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{PCA}} - \boldsymbol{\Sigma})\|_2)}{\min_{j\neq i}|\lambda_j - \lambda_i|} \lesssim p^{-1/2}T^{-1/2} + p^{-1}.$$

Moreover, if $\widehat{\boldsymbol{w}}_{i,\mathrm{PCA}}^\top\boldsymbol{w}_i \ge 0$, *then* $\mathbb{E}(\|\widehat{\boldsymbol{w}}_{i,\mathrm{PCA}} - \boldsymbol{w}_i\|_2) \lesssim p^{-1/2}T^{-1/2} + p^{-1}$ *for each* $i = 1, \ldots, K$.

Next, analogous to Theorem 3.2, we will show the approximation error rate to the distribution of the standardized deviation between $\boldsymbol{w}_i$ and $\widehat{\boldsymbol{w}}_i$ by the standard normal distribution, namely the Berry-Esseen type bound. Denote $\mathbf{P}_i = \boldsymbol{w}_i(\boldsymbol{w}_i'\boldsymbol{w}_i)^{-1}\boldsymbol{w}_i'$ and $\widehat{\mathbf{P}}_i = \widehat{\boldsymbol{w}}_i(\widehat{\boldsymbol{w}}_i'\widehat{\boldsymbol{w}}_i)^{-1}\widehat{\boldsymbol{w}}_i'$ the projectors onto the spaces spanned by $\boldsymbol{w}_i$ and $\widehat{\boldsymbol{w}}_i$ respectively, we first establish the Berry-Esseen type bound for $\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_{\mathbb{F}}^2$ below.

Theorem 4.6. *Under Conditions 2.1-2.3, given* $\mathbf{Y}$ *from (1.1) or (1.2), for each* $i = 1, \ldots, p$,

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_{\mathbb{F}}^2 - \mathbb{E}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_{\mathbb{F}}^2)}{\mathrm{Var}^{1/2}(\|\widehat{\mathbf{P}}_i - \mathbf{P}_i\|_{\mathbb{F}}^2)} \le x\right\} - \Phi(x)\right| \lesssim \frac{1}{B_i} + \frac{\log T}{\sqrt{T}} + \frac{(\log T)^{1/2}(\log p)^{1/4}}{T^{1/8}B_i},$$

where $B_i = 2\sqrt{2}\|\mathbf{P}_i\boldsymbol{\Sigma}\mathbf{P}_i\|_{\mathbb{F}}\|\mathbf{Q}_i\boldsymbol{\Sigma}\mathbf{Q}_i\|_{\mathbb{F}}$ *and* $\mathbf{Q}_i = \sum_{j\neq i}(\lambda_i - \lambda_j)^{-1}\mathbf{P}_j$.

A similar result has been documented for independent data in literature (Koltchinskii and Lounici [68]); while, Theorem 4.6 is more general by allowing temporal dependence in data. In fact, the third term on the right hand side above quantifies the effect of temporal dependence, and as a result, the convergence rate is slightly compromised compared to the rate under independence. As the Frobenius norm and ℓ_2-norm of a vector are the same, Theorem 4.6 leads to the following corollary, which extends the Berry-Esseen bound for random vectors (Goldstein and Shao [52], Bobkov and Chistyakov [26], Bobkov, Chistyakov, and Götze [25]).

Corollary 4.3. *Under the same conditions in Theorem 4.6, for any matrix* $\mathbf{C}$ *and* $i = 1, \ldots, p$,

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\|\widehat{\mathbf{P}}_i\mathbf{C} - \mathbf{P}_i\mathbf{C}\|_{\mathbb{F}}^2 - \mathbb{E}(\|\widehat{\mathbf{P}}_i\mathbf{C} - \mathbf{P}_i\mathbf{C}\|_{\mathbb{F}}^2)}{\mathrm{Var}^{1/2}(\|\widehat{\mathbf{P}}_i\mathbf{C} - \mathbf{P}_i\mathbf{C}\|_{\mathbb{F}}^2)} \le x\right\} - \Phi(x)\right| \lesssim \frac{1}{B_i} + \frac{\log T}{\sqrt{T}} + \frac{(\log T)^{1/2}(\log p)^{1/4}}{T^{1/8}B_i}.$$

Particularly, for each $i = 1, \ldots, p$,

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2 - \mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2)}{\mathrm{Var}^{1/2}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2)} \le x\right\} - \Phi(x)\right| \lesssim \frac{1}{B_i} + \frac{\log T}{\sqrt{T}} + \frac{(\log T)^{1/2}(\log p)^{1/4}}{T^{1/8}B_i}.$$

Note that $B_i = O(\sqrt{p})$ for $i = 1, \ldots, K$. Thus, Corollary 4.3 provides a uniform normal approximation to standardized $\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2$ for $i = 1, \ldots, K$. However, $B_i = O(1)$ for $i > K$ so that the upper bounds in both Theorem 4.6 and Corollary 4.3 do not necessarily shrink to zero. Therefore, as noted by Koltchinskii and Lounici [67], the normal approximation to $\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2$ for $i > K$ may fail to hold. Together with Theorem 3.3, Corollary 4.3 shows that, the spectral structures corresponding to the spiked eigenvalues, *i.e.*, the first K eigenvalues of the sample covariance matrix, provide good estimates to the corresponding spectral structures of $\boldsymbol{\Sigma}$, even for $p > T$ for which $\widehat{\boldsymbol{\Sigma}}$ is no longer consistent to $\boldsymbol{\Sigma}$.

Remark 4.1. In practice, $\mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2)$ and $\mathrm{Var}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2)$ are unknown. To use Corollary 4.3 for inference, we need to estimate them. Koltchinskii and Lounici [68] offered a data-splitting procedure which splits the sample into three subsamples: the first for estimating the expectation, the second for estimating the variance, and the third for building the confidence set. In addition, since $T^{-1}\mathbf{Y}\mathbf{Y}^\top$ is naturally an empirical process, the multiplier bootstrap can be employed to build the confidence set of $\boldsymbol{w}_i$ for each $i = 1, \ldots, K$ without data splitting for *i.i.d* data (Naumov, Spokoiny, and Ulyanov [74]). Under Condition 2.3, $\boldsymbol{y}_t$ from (1.1) is weakly temporal dependent and can be approximated by some m-dependent time series $\widetilde{\boldsymbol{y}}_t$ in the following sense,

$$|\mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2|\boldsymbol{y}_t) - \mathbb{E}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2|\widetilde{\boldsymbol{y}}_t)| \lesssim \frac{(\log T)^{1/2}(\log p)^{1/4}}{T^{9/8}},$$

$$|\,\mathrm{Var}^{1/2}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2|\boldsymbol{y}_t) - \mathrm{Var}^{1/2}(\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2^2|\widetilde{\boldsymbol{y}}_t)| \lesssim \frac{(\log T)^{1/2}(\log p)^{1/4}}{T^{9/8}};$$

and $\|\widehat{w}_i - w_i\|_2^2$ based on $\boldsymbol{y}_t$ and $\widetilde{\boldsymbol{y}}$ have the similar normal approximations (Chen and Shao [37,38], Zhang and Cheng [94]). Therefore, we can employ the following blockwise multiplier bootstrap procedure to draw inference on $\boldsymbol{w}_i$ (Zhang and Cheng [94]), whose guarantee is provided by Corollary 4.3 and the above approximation using $\widetilde{\boldsymbol{y}}_t$.

Algorithm: Blockwise multiplier bootstrap procedure for the inference of $\boldsymbol{w}_i$

Input: Observations $\{y_{it}\}_{i=1,t=1}^{p,T}$.

Step 1. Pre-specify integers b_T and l_T such that $T = b_T l_T$ based on the nonparametric plug-in method (Bühlmann and Künsch [28]), the empirical criteria-based method (Hall, Horowitz, and Jing [53]) or the algorithm in Zhang and Cheng [94].

Step 2. Generate e_{js} *i.i.d.* from $\mathcal{N}(1, 1)$ for $j = 1, \ldots, B$ and $s = 1, \ldots, l_T$.

Step 3. For each j, calculate $\boldsymbol{\Sigma}_j^{\mathrm{BS}} = T^{-1}\sum_{s=1}^{l_T} e_{js} \sum_{t=(s-1)b_T+1}^{sb_T} \boldsymbol{y}_t \boldsymbol{y}_t'$.

Step 4. For each $i = 1, \ldots, K$, denote $\boldsymbol{w}_{i,j}^{\mathrm{BS}}$ the eigenvector corresponding to the ith largest eigenvalue of $\boldsymbol{\Sigma}_j^{\mathrm{BS}}$ and define $\gamma_\alpha^{\mathrm{BS}}$ as the $1-\alpha$ percentile of $\{\|\boldsymbol{w}_{i,j}^{\mathrm{BS}} - \widehat{\boldsymbol{w}}_i\|_2^2\}_{j=1}^B$.

Output: Confidence set of $\boldsymbol{w}_i$ as $\{\boldsymbol{w} : \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_i\|_2^2 \le \gamma_\alpha^{\mathrm{BS}}\}$ for $i = 1, \ldots, K$.

4.4. Low-rank matrix denoising based on temporally dependent data

Low-rank matrix denoising has numerous applications such as robust video restoration (Ji et al. [60]), hyperspectral image restoration (He et al. [54], Zhang et al. [92]), and underdetermined direction of arrival estimation (Pal and Vaidyanathan [77]). Lately, the low-rank matrix denoising in the presence of both heteroskedastic errors and dependent samples have attracted great attention in literature (Zhang, Cai, and Wu [91]). Suppose we observe time series

$$y_{it} = x_{it} + u_{it}$$

for $i = 1, \ldots, p$ and $t = 1, \ldots, T$, which can be written as

$$\mathbf{Y} = \mathbf{X} + \mathbf{U}$$

where $\mathbf{Y} = \{y_{it}\}_{i=1,t=1}^{p,T}$, $\mathbf{X} = \{x_{it}\}_{i=1,t=1}^{p,T}$ is a fixed rank-K matrix, and $\mathbf{U} = \{u_{it}\}_{i=1,t=1}^{p,T}$. Assume the noise matrix $\mathbf{U}$ satisfies Condition 2.3. Let $\mathbf{X} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{V}'$ be the SVD, where $\mathbf{W}$ is a $p \times K$ orthogonal

matrix and $\mathbf{V}$ is a $T \times K$ orthogonal matrix. Note that the column space of $\mathbf{W}$ is essentially that of $\mathbf{A}$ in (1.2) under Condition 2.1. Then we can use PCA to estimate $\mathbf{W}$ by $\widehat{\mathbf{W}} = (\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}})^{-1/2}\widehat{\mathbf{A}}$ with the following theoretical guarantees.

Corollary 4.4. *Suppose that* $p \lesssim \lambda_{\min}(\mathbf{\Lambda}) \lesssim \lambda_{\max}(\mathbf{\Lambda}) \lesssim p$. *Then* $\mathbf{W}$ *and* $\widehat{\mathbf{W}}$ *satisfy*

$$\mathbb{E}\{\|\sin\Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_{\mathbb{F}}\} \lesssim \frac{1}{\sqrt{p}} + \frac{1}{\sqrt{T}},$$

where $\|\sin\Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_{\mathbb{F}} \overset{d}{=} \|\mathbf{W}_\perp^\top \widehat{\mathbf{W}}\|_{\mathbb{F}}$ *and* $\mathbf{W}_\perp$ *is a* $p \times (p-K)$ *orthogonal matrix such that* $(\mathbf{W}, \mathbf{W}_\perp)$ *is a* $p \times p$ *orthogonal matrix.*

In Corollary 4.4, we consider a spike model with potentially heteroskedastic errors. Like the approximate factor model, the spiked singular values of $\mathbf{X}$ provide stronger signals compared to the model used in traditional matrix denoising (Cai and Zhang [32], Zhang, Cai, and Wu [91]). To compare, for the non-spiked signal matrix $\mathbf{X}$ and homoskedastic variance of $\mathbf{U}$, the optimal rate of matrix denoising using the regular SVD is $\mathbb{E}(\|\sin\Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_{\mathbb{F}}) \lesssim \min(p, T)^{-1/2}$ (Theorems 3 and 4, Cai and Zhang [32]). Thus, Corollary 4.4 gives similar results to the regular SVD (Cai and Zhang [32]) and the diagonal-deletion SVD (Florescu and Perkins [51]). In addition, Theorem 4 in Zhang, Cai, and Wu [91] showed that the heteroskedastic PCA can obtain the optimal rate of matrix denoising for non-spiked signal matrix $\mathbf{X}$ with heteroskedastic errors. It is easy to see that if the variance of u_{it} is bounded for each i and t, the optimal rate in Zhang, Cai, and Wu [91] is also $\mathbb{E}(\|\sin\Theta(\widehat{\mathbf{W}}, \mathbf{W})\|_{\mathbb{F}}) \lesssim \min(p, T)^{-1/2}$. Hence, our result also matches the heteroskedastic PCA (Zhang, Cai, and Wu [91]) in the presence of heteroskedastic errors.

5. Numerical studies

In this section, we perform simulation studies to further illustrate results displayed in Sections 3, 4.1, and 4.2.

We first conduct numerical experiments to demonstrate Theorem 3.3. Consider model (1.1) with $a_{i1}, \ldots, a_{p1} = 1$ for $K = 1$, $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$, and three settings for one-dimensional latent process f_{t1}: (1) AR(1) with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 0.75)$ innovation; (2) AR(1) with autoregressive coefficient $\phi = 0.5$ and $t_8/\sqrt{0.75}$ innovation; and (3) ARMA(1, 1) with autoregressive coefficient $\phi = 0.5$, moving average coefficient $\theta = 0.5$, and $\mathcal{N}(0, 3/7)$ innovation. Under these settings, $\lambda_1 = p + 0.01$ while other eigenvalues, such as λ_2 and λ_{10}, are all equal to 0.01 for any p. For the ease of visualization, the variance of u_{it} is particularly set to be 0.01 so that the error process does not affect much on estimating the eigenvalues of the covariance matrix. In fact, the simulation results, especially the trends of errors versus p or T, are not sensitive to the variance of u_{it}. This can be seen from the results of extra numerical experiments, where $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ so that the variance of u_{it} is enlarged 100 times and f_{t1} follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 0.75)$ innovation (see Figure S.1 in the Supplementary Material (Zhang, Zhou, and Wang [93])). Two scenarios on p and T are considered, $p = \lfloor 2T^{1/2} \rfloor$ and $T = \lfloor 2p^{1/2} \rfloor$. Based on 100 replicates, the simulation results are displayed in Figure 1. From panels (a1), (b1), and (c1), we can see that $\widehat{\lambda}_i$ converges to λ_i when $p < T$. The relative error $|\widehat{\lambda}_i/\lambda_i - 1|$ for $i = 1$ converges to zero faster than those for $i = 2$ and 10 since λ_1 diverges in p while λ_2 and λ_{10} remain in constants. In addition, from panels (a2), (b2), and (c2), it is noticed that $\widehat{\lambda}_1$ still converges to λ_1 even for $p > T$ while the deviations

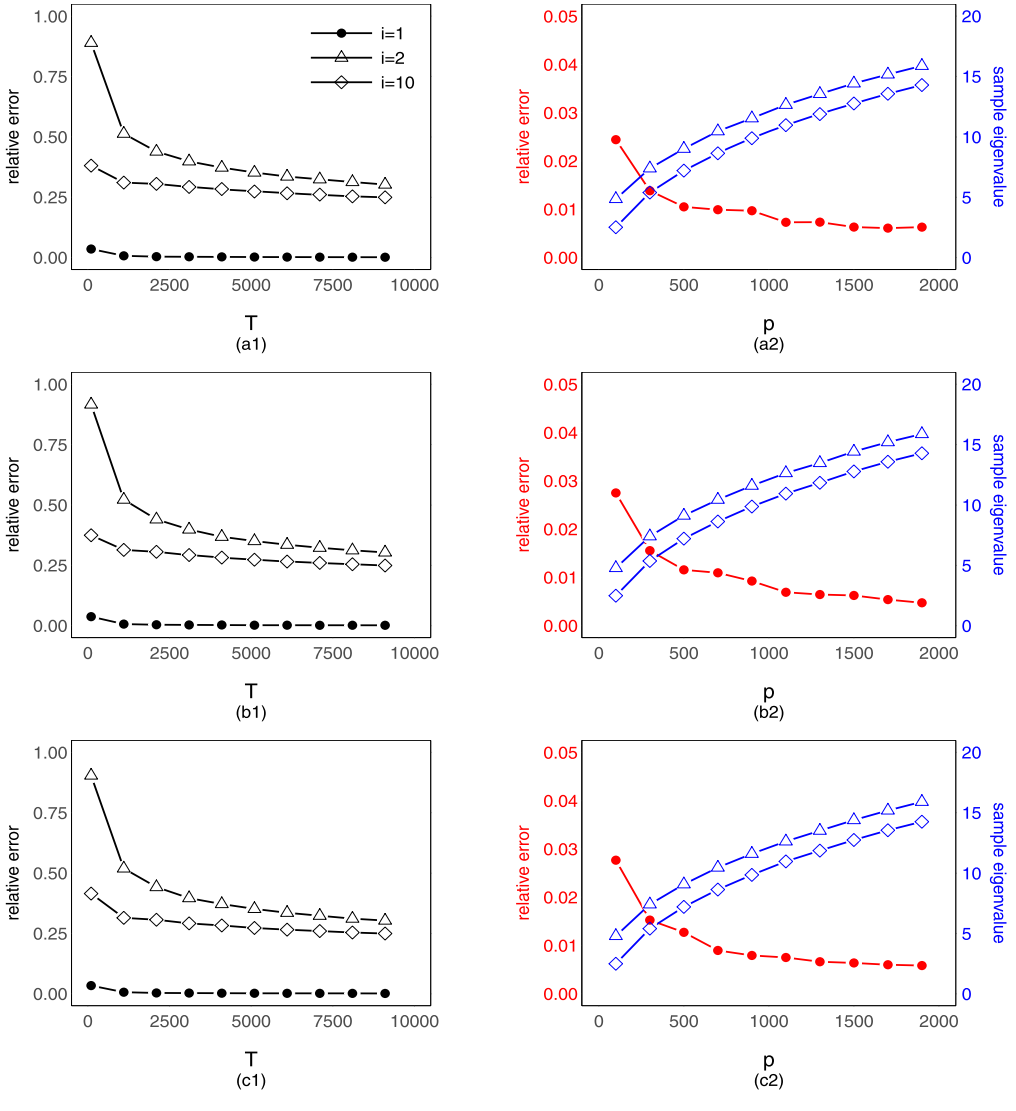

Figure 1. In the simulation, $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$. In the left column, $p = \lfloor 2T^{1/2} \rfloor$ ($p < T$), and in the right column, $T = \lfloor 2p^{1/2} \rfloor$ ($p > T$). In panels (a1) and (a2), latent process f_{t1} follows setting (1); in panels (b1) and (b2), latent process f_{t1} follows setting (2); and in panels (c1) and (c2), latent process f_{t1} follows setting (3). In panels (a1), (b1), and (c1), the relative errors $|\widehat{\lambda}_i/\lambda_i - 1|$ for $i = 1, 2, 10$ are displayed. In panels (a2), (b2), and (c2), the relative errors are displayed for λ_1 and the sample eigenvalues are displayed for λ_2 and λ_{10} to show that they are unbounded in p.

of other eigenvalues diverge as p and T diverge. These patterns are commonly observed for all three settings on f_{t1}. This matches results in Theorem 3.3.

Next, we demonstrate the influence of p, T, and eigenvalues of $\mathbf{\Sigma}$ on the probability of estimating the correct number of factors using the ratio of consecutive eigenvalues in (4.1). Consider model (1.1) with $K = 3$ factors and $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 25)$. The three components in $\mathbf{f}_t = (f_{t1}, f_{t2}, f_{t3})^\top$ are in-

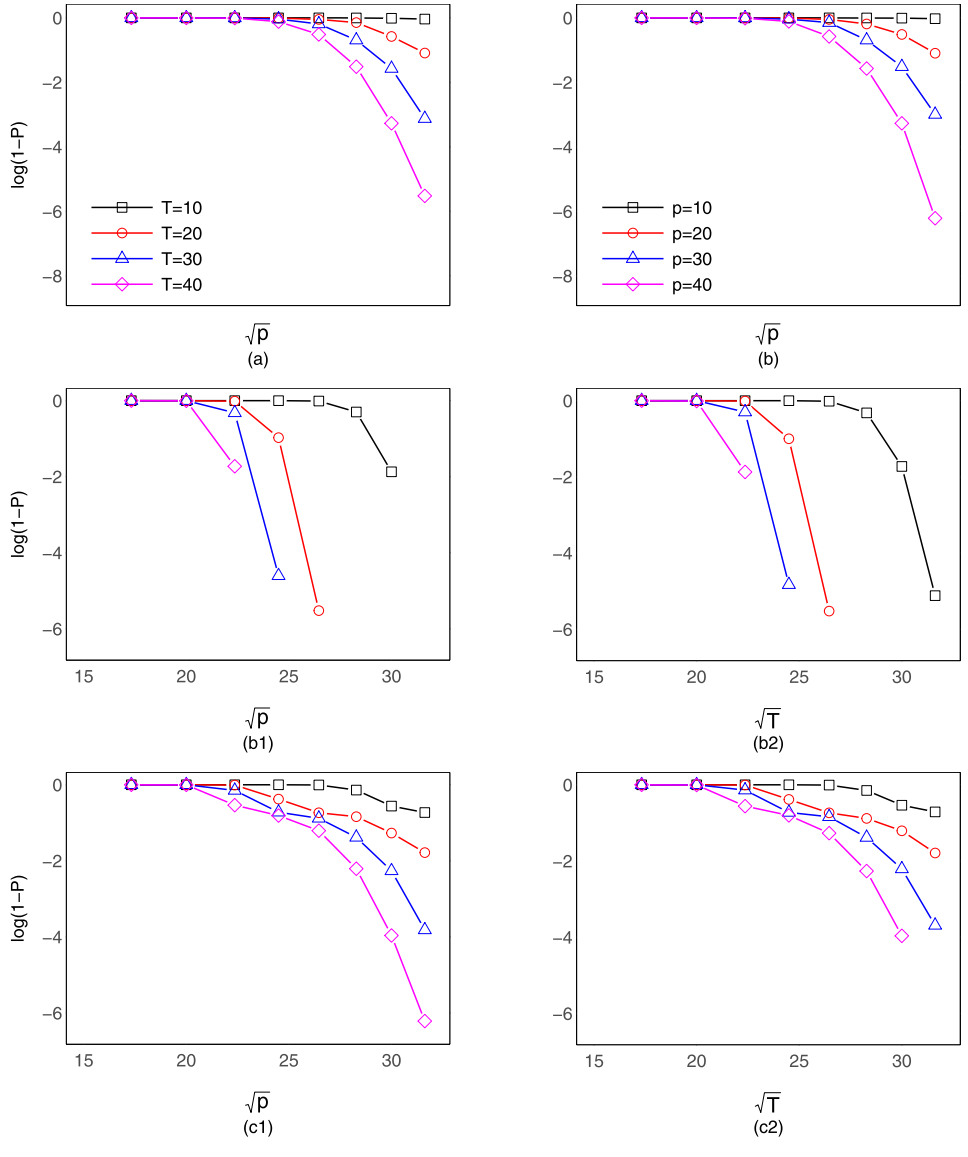

Figure 2. Plots about $\log(1-\mathbb{P}\{\widehat{K}=K\})$ for $p=100,200,\ldots,1000$, $T=500,700,800,900$ (left column), and $T=100,200,\ldots,1000$, $p=500,700,800,900$ (right column). The diagonal entries in $p^{-1}\mathbf{A}^{\top}\mathbf{A}$ are $\{16,4,1\}$ (panels (a1) and (a2)), $\{16,4,2\}$ (panels (b1) and (b2)), and $\{32,4,2\}$ (panels (c1) and (c2)). Points are omitted when $\log(1-\mathbb{P}\{\widehat{K}=K\})=-\infty$, *i.e.*, $\mathbb{P}(\widehat{K}=K)=1$.

dependent and identical AR(1) processes with autoregressive coefficients $\phi=0.5$ and $\mathcal{N}(0,0.75)$ innovation. We further set $\mathbf{A}$ such that $p^{-1}\mathbf{A}^{\top}\mathbf{A}$ has diagonal entries $\{16,4,1\}$ (panels (a1) and (a2) in Figure 2), $\{16,4,2\}$ (panels (b1) and (b2) in Figure 2), and $\{32,4,2\}$ (panels (c1) and (c2) in Figure 2). For p and T, two settings are reported: (1) T is fixed, $p=100,200,\ldots,1000$; and (2) p is fixed, $T=100,200,\ldots,1000$. Based on 500 replicates, results on $\log(1-\mathbb{P}\{\widehat{K}=K\})$ are displayed in Figure 2. In practice, the method using ratios of consecutive eigenvalues performs well to estimate the

number of latent factors. For instance, under the setting that $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, three components in $\mathbf{f}_t$ are independent and identical AR(1) processes with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 0.75)$ innovation, and the diagonal entries of $p^{-1}\mathbf{A}^\top\mathbf{A}$ are $\{16, 4, 2\}$ in model (1.1), $\mathbb{P}(\widehat{K} = K)$ quickly approaches 1 even for relatively small p and T (see Figure S.2 in the Supplementary Material (Zhang, Zhou, and Wang [93])). Thus, for the ease of visualization, here we set the variance of u_{it} large so that the trend of estimation errors on K versus p and T can be displayed clearly. In Figure 2, we notice that $\log(1 - \mathbb{P}\{\widehat{K} = K\})$ decreases faster for greater λ_K/λ_{K+1} and smaller $\max_{i\neq K} \lambda_i/\lambda_{i+1}$. In fact, from Theorem 4.1, $\log(1 - \mathbb{P}\{\widehat{K} = K\})$ is bounded by a quadratic function of $\sqrt{\max(p, T)}$ with C_1 and C_2 defined in Theorem 4.1. Since c and C in Theorem 3.3 only depend on the distribution of $\boldsymbol{u}_t$, C_2 is the same for different $\mathbf{A}$. On the other hand, as λ_K/λ_{K+1} increases and $\max_{i\neq K} \lambda_i/\lambda_{i+1}$ decreases, C_1 increases so that the quadratic function of $\sqrt{\max(p, T)}$ has a smaller vertex and greater quadratic coefficient. Thus, Figure 2 confirms the conclusion in Theorem 4.1.

Finally, we study the estimation of moments of latent factor process $\mathbf{f}_t$ to demonstrate Theorem 4.4. Still consider model (1.1) with $K = 1$ factor and $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$. Also, we set three models for the univariate latent factor process f_{t1}: (1) AR(1) with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 0.75)$ innovation; (2) AR(1) with autoregressive coefficient $\phi = 0.5$ and $t_8/\sqrt{0.75}$ innovation; and (3) ARMA(1, 1) with autoregressive coefficient $\phi = 0.5$, moving average coefficient $\theta = 0.5$ and $\mathcal{N}(0, 3/7)$ innovation. Two settings about p and T are considered: $p = 200$ with $T = 100, 200, \ldots, 1000$; and $T = 200$ with $p = 100, 200, \ldots, 1000$. Based on 100 replicates, $|\widehat{\rho}(h, \widehat{f}_{t1}) - \widehat{\rho}(h, f_{t1})|$ and $|\widehat{\Psi}(h, \widehat{f}_{t1}) - \widehat{\Psi}(h, f_{t1})|$ versus T and p are displayed in log-log scale in Figures 3 and 4. For all settings, the squared differences for both ACF and PACF shrink to zero as p and T diverge. Also, in all settings, the slopes of the log difference of ACF or PACF versus $\log T$ or $\log p$ are $-1/2$ (the red lines), which confirms the established rates of convergence in Theorem 4.4.

6. Conclusions

In this paper, we scrupulously study the non-asymptotic properties of the spectral decomposition of large Gram-type matrices under the assumption that the data matrix $\mathbf{Y}$ is governed by a factor model. As a result, we establish the exponential tail bound for the first and second moments of the deviation between the empirical and population eigenvectors to the right Gram matrix as well as the Berry-Esseen type bound to characterize the Gaussian approximation of these deviations. Technically, we successfully relax the assumption upon latent factors in the factor model, so that the latent factor processes are no longer restricted to a subspace as stated by Condition PC1 in Bai and Ng [13]. We also obtain the non-asymptotic tail bound of the ratio between eigenvalues of the sample covariance matrix, and their population counterparts regardless of the size of the data matrix. This extends the works of Bai and Yin [16], Lam and Yao [69], and Wang and Fan [88].

With the derived non-asymptotic properties of eigenvalues of the sample covariance matrix, we provide the non-asymptotic characterization of different consecutive-eigenval-ues-based methods to estimate the number of latent factors in factor models and relate machine learning problems. The established non-asymptotic lower bound of the probability of estimating the correct number of factors reveal the influence of p, T and eigenvalues of $\boldsymbol{\Sigma}$ on different methods. In addition, as an application of our main results, we provide statistical guarantees on estimating the parametric models for the latent process in dynamic or approximate factor models, so that one can make forecast based on the factor models and high-dimensional time series. We also obtain non-asymptotic properties of the spectral structure of large sample covariance matrices, including the Davis-Kahan type perturbation result and the approximation error rate to the distribution of the standardized deviation between $\boldsymbol{w}_i$ and $\widehat{\boldsymbol{w}}_i$ by the standard normal distribution, *i.e.* the Berry-Esseen type bound. Based on these results, it is possible

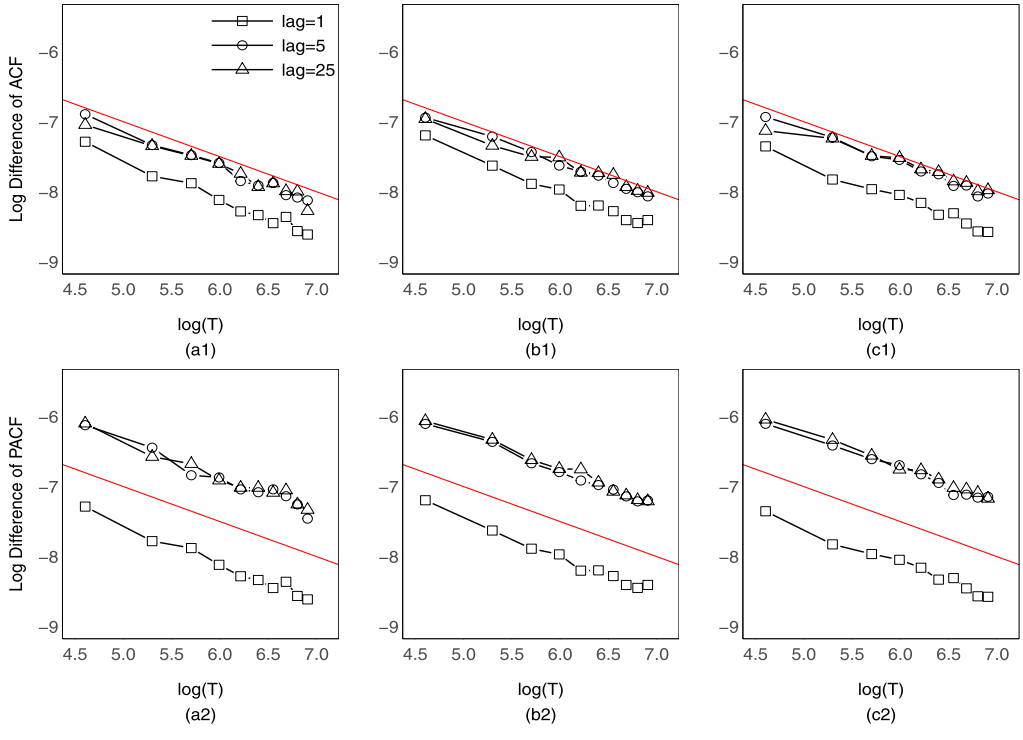

Figure 3. Log differences of ACF (first row) and PACF (second row) of $\{f_{t1} : t \geq 1\}$ at lag $h = 1$, lag $h = 5$, and lag $h = 25$ for $p = 200$ and $T = 100, 200, \ldots, 1000$. The latent process follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 1)$ innovation in panels (a1) and (a2); it follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and t_8 innovation in panels (b1) and (b2); and it follows ARMA(1, 1) with autoregressive coefficient $\phi = 0.5$, moving average coefficient $\theta = 0.5$, and $\mathcal{N}(0, 1)$ innovation in panels (c1) and (c2). The red solid line has slope $-1/2$.

to construct confidence sets for the leading eigenvectors of $\mathbf{\Sigma}$ using the multiplier bootstrap. Finally, we apply our results to the low-rank matrix denoising in the presence of heteroskedastic errors and temporal dependence in data.

Acknowledgments

The authors thank the Editor, an Associate Editor, and a reviewer for many helpful and constructive comments. The work of Wen Zhou was partially supported by Department of Energy grant DE-SC0018344 and National Science Foundation grants IIS-1545994 and IOS-1922701. The research of Haonan Wang was partially supported by National Science Foundation grants DMS-1737795, DMS-1923142 and CNS-1932413.

Supplementary Material

Additional proofs and numerical results (DOI: 10.3150/21-BEJ1384SUPP; .pdf). The Supplementary Material contains technical results used for the main paper. In Section A, we prove the main results

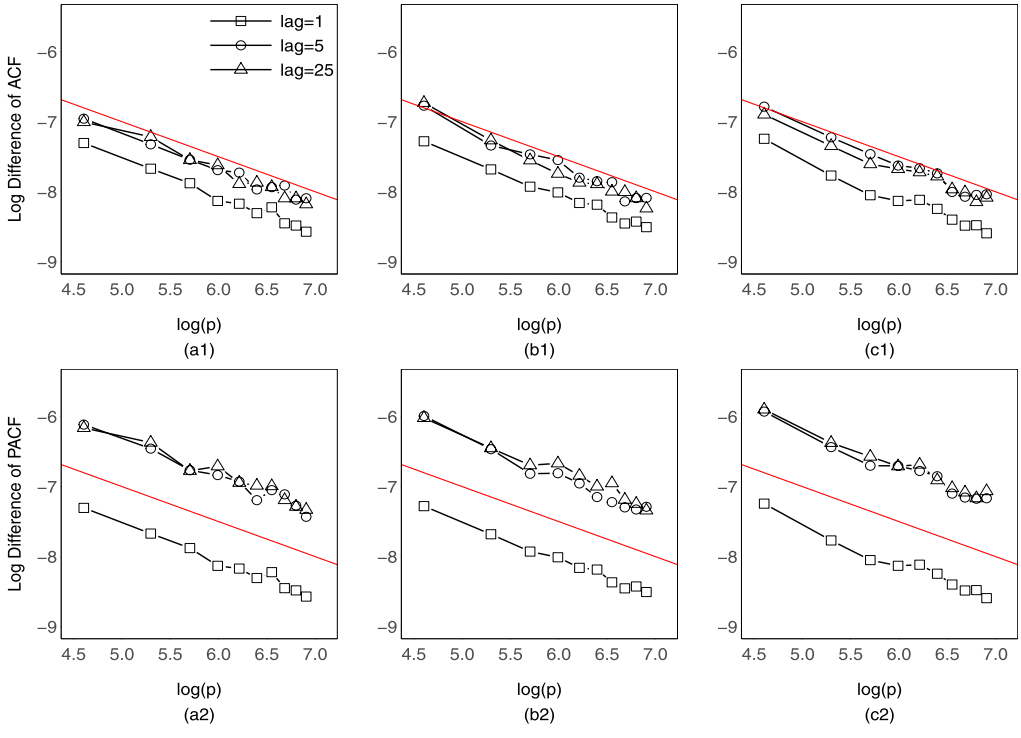

Figure 4. Log differences of ACF (first row) and PACF (second row) of $\{f_{t1} : t \geq 1\}$ at lag $h = 1$, lag $h = 5$, and lag $h = 25$ for $T = 200$ and $p = 100, 200, \ldots, 1000$. The latent process follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and $\mathcal{N}(0, 1)$ innovation in panels (a1) and (a2); it follows AR(1) process with autoregressive coefficient $\phi = 0.5$ and t_8 innovation in panels (b1) and (b2); and it follows ARMA(1, 1) with autoregressive coefficient $\phi = 0.5$, moving average coefficient $\theta = 0.5$, and $\mathcal{N}(0, 1)$ innovation in panels (c1) and (c2). The red solid line has slope $-1/2$.

in Theorems 3.1-3.3. In Section B, we show Theorems 4.1-4.6. Section C includes technical lemmas and auxiliary results, and extra numerical results are reported in D.

References

[1] Adamczak, R., Litvak, A., Pajor, A. and Tomczak-Jaegermann, N. (2010). Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *J. Amer. Math. Soc.* **23** 535–561.

[2] Ahn, S.C. and Horenstein, A.R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227.

[3] Ahn, S.C., Lee, Y.H. and Schmidt, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. *J. Econometrics* **101** 219–255.

[4] Anderson, T.W. (1962). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

[5] Anderson, T.W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Stat.* **34** 122–148.

[6] Anderson, T.W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume* 5: *Contributions to Econometrics, Industrial Research, and Psychometry* 111–150. Univ. California Press.

[7] Athreya, K.B. and Lahiri, S.N. (2006). *Measure Theory and Probability Theory*. New York: Springer.

[8] Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171.
[9] Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229–1279.
[10] Bai, J. and Li, K. (2014). Theory and methods of panel data models with interactive effects. *Ann. Statist.* **42** 142–170.
[11] Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.
[12] Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Found Trends Econom.* **3** 89–163.
[13] Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *J. Econometrics* **176** 18–29.
[14] Bai, Z., Choi, K.P. and Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *Ann. Statist.* **46** 1050–1076.
[15] Bai, Z. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. New York: Springer.
[16] Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.* **21** 1275–1294.
[17] Bai, Z., Yin, Y. and Krishnaiah, P.R. (1986). On limiting spectral distribution of product of two random matrices when the underlying distribution is isotropic. *J. Multivariate Anal.* **19** 189–200.
[18] Bai, Z., Yin, Y. and Krishnaiah, P.R. (1988). On the limiting empirical distribution function of the eigenvalues of a multivariate F matrix. *Theory Probab. Appl.* **32** 490–500.
[19] Baik, J., Arous, G.B. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697.
[20] Bartholomew, D.J., Knott, M. and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis*: *A Unified Approach*. New York: Wiley.
[21] Bialecki, B. and Fairweather, G. (1995). Matrix decomposition algorithms in orthogonal spline collocation for separable elliptic boundary value problems. *SIAM J. Sci. Comput.* **16** 330–347.
[22] Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604.
[23] Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227.
[24] Bien, J., Bunea, F. and Xiao, L. (2016). Convex banding of the covariance matrix. *J. Amer. Statist. Assoc.* **111** 834–845. https://doi.org/10.1080/01621459.2015.1058265
[25] Bobkov, S.G., Chistyakov, G. and Götze, F. (2018). Berry–Esseen bounds for typical weighted sums. *Electron. J. Probab.* **23**.
[26] Bobkov, S.G. and Chistyakov, G.P. (2015). On concentration functions of random variables. *J. Theoret. Probab.* **28** 976–988.
[27] Brockwell, P.J., Davis, R.A. and Fienberg, S.E. (1991). *Time Series*: *Theory and Methods*: *Theory and Methods*. New York: Springer.
[28] Bühlmann, P. and Künsch, H.R. (1999). Block length selection in the bootstrap for time series. *Comput. Statist. Data Anal.* **31** 295–310.
[29] Bunea, F. and Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli* **21** 1200–1230.
[30] Cai, T., Han, X. and Pan, G. (2020). Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *Ann. Statist.* **48** 1255–1280.
[31] Cai, T., Liu, W. and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277.
[32] Cai, T. and Zhang, A. (2016). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *J. Multivariate Anal.* **150** 55–74.
[33] Callaert, H. and Janssen, P. (1978). The Berry-Esseen theorem for U-statistics. *Ann. Statist.* **6** 417–421.
[34] Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304.
[35] Chan, Y.-K. and Wierman, J. (1977). On the Berry-Esseen theorem for U-statistics. *Ann. Probab.* **5** 136–139.
[36] Chen, L., Wang, W. and Wu, W.B. (2021). Dynamic semiparametric factor model with structural breaks. *J. Bus. Econom. Statist.* **39** 757–771. MR4272933 https://doi.org/10.1080/07350015.2020.1730857

[37] Chen, L.H. and Shao, Q.-M. (2004). Normal approximation under local dependence. *Ann. Probab.* **32** 1985–2028.
[38] Chen, L.H. and Shao, Q.-M. (2007). Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli* **13** 581–599.
[39] Chen, X., Womersley, R.S. and Ye, J.J. (2011). Minimizing the condition number of a Gram matrix. *SIAM J. Optim.* **21** 127–148.
[40] Choi, Y., Taylor, J. and Tibshirani, R. (2017). Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *Ann. Statist.* **45** 2590–2617.
[41] Davis, C. and Kahan, W.M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46.
[42] De Almeida, M.C., Asada, E.N. and Garcia, A.V. (2008). On the use of gram matrix in observability analysis. *IEEE Trans. Power Syst.* **23** 249–251.
[43] De Almeida, M.C., Asada, E.N. and Garcia, A.V. (2008). Power system observability analysis based on gram matrix and minimum norm solution. *IEEE Trans. Power Syst.* **23** 1611–1618.
[44] Donath, W.E. and Hoffman, A.J. (1973). Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.* **17** 420–425.
[45] Drineas, P. and Mahoney, M.W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.* **6** 2153–2175.
[46] Fan, J., Ke, Y., Sun, Q. and Zhou, W.-X. (2019). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *J. Amer. Statist. Assoc.* **114** 1880–1893.
[47] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. Roy. Statist. Soc. Ser. B* **75** 603–680.
[48] Fan, J., Liao, Y. and Wang, W. (2016). Projected principal component analysis in factor models. *Ann. Statist.* **44** 219–254.
[49] Fan, J., Sun, Q., Zhou, W. and Zhu, Z. (2018). Principal component analysis for big data. Wiley StatsRef: Statistics Reference Online. to appear. https://doi.org/10.1002/9781118445112.stat08122
[50] Fan, J., Wang, W. and Zhong, Y. (2018). An ℓ^∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** 1–42.
[51] Florescu, L. and Perkins, W. (2016). Spectral thresholds in the bipartite stochastic block model. *Proc. Mach. Learn. Res.* **49** 943–959.
[52] Goldstein, L. and Shao, Q.-M. (2009). Berry-Esseen bounds for projections of coordinate symmetric random vectors. *Electron. Commun. Probab.* **14** 474–485.
[53] Hall, P., Horowitz, J.L. and Jing, B. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika* **82** 561–574.
[54] He, W., Zhang, H., Zhang, L. and Shen, H. (2015). Hyperspectral image denoising via noise-adjusted iterative low-rank matrix approximation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **8** 3050–3061.
[55] Hörmann, S. (2009). Berry-Esseen bounds for econometric time series. *ALEA Lat. Am. J. Probab. Math. Stat.* **6** 377–397.
[56] Horst, P. (1965). *Factor Analysis of Data Matrices*. New York: Holt, Rinehart and Winston.
[57] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24** 417.
[58] James, G. and Murphy, G. (1979). The determinant of the gram matrix for a Specht module. *J. Algebra* **59** 222–235.
[59] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume* 1: *Contributions to the Theory of Statistics*. The Regents of the Univ. California.
[60] Ji, H., Huang, S., Shen, Z. and Xu, Y. (2011). Robust video restoration by joint sparse and low rank matrix approximation. *SIAM J. Imaging Sci.* **4** 1122–1142.
[61] Jirak, M. (2016). Berry–Esseen theorems under weak dependence. *Ann. Probab.* **44** 2024–2063.
[62] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327.
[63] Johnstone, I.M. and Lu, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.

[64] Johnstone, I.M. and Paul, D. (2018). PCA in high dimensions: An orientation. *Proc. IEEE* **106** 1277–1292.
[65] Jolliffe, I. (2002). *Principal Component Analysis*, 2nd ed. ed. New York: Springer.
[66] Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.* **12** 1–38.
[67] Koltchinskii, V. and Lounici, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.* **52** 1976–2013.
[68] Koltchinskii, V. and Lounici, K. (2017). Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.* **45** 121–157.
[69] Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726.
[70] Lawley, D.N. and Maxwell, A.E. (1962). Factor analysis as a statistical method. *J. R. Stat. Soc., Ser. D* **12** 209–229.
[71] Mandel, J. (1982). Use of the singular value decomposition in regression analysis. *Amer. Statist.* **36** 15–24.
[72] Merlevède, F., Peligrad, M. and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probab. Theory Related Fields* **151** 435–474.
[73] Moon, H.R. and Weidner, M. (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory* **33** 158–195.
[74] Naumov, A., Spokoiny, V. and Ulyanov, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probab. Theory Related Fields* **174** 1091–1132.
[75] Ng, A.Y., Jordan, M.I. and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 849–856.
[76] Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics* **168** 244–258.
[77] Pal, P. and Vaidyanathan, P.P. (2014). A grid-less approach to underdetermined direction of arrival estimation via low rank matrix denoising. *IEEE Signal Process. Lett.* **21** 737–741.
[78] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2** 559–572.
[79] Ramona, M., Richard, G. and David, B. (2012). Multiclass feature selection with kernel Gram-matrix-based criteria. *IEEE Trans. Neural Netw. Learn. Syst.* **23** 1611–1623.
[80] Rummel, R.J. (1988). *Applied Factor Analysis*. Evanston, IL: Northwestern Univ. Press.
[81] Schölkopf, B., Shawe-Taylor, J., Smola, A.J. and Williamson, R.C. (1999). Generalization bounds via eigenvalues of the Gram matrix Technical Report 99-035 NeuroCOLT.
[82] Shawe-Taylor, J., Williams, C., Cristianini, N. and Kandola, J. (2002). On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In *International Conference on Algorithmic Learning Theory* 23–40. Springer.
[83] Shawe-Taylor, J., Williams, C.K., Cristianini, N. and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inf. Theory* **51** 2510–2522.
[84] Stark, C. (2014). Self-consistent tomography of the state-measurement Gram matrix. *Phys. Rev. A* **89** 052109.
[85] Stein, C. (1956). Some problems in multivariate analysis, Part I Technical report, Stanford Univ.
[86] Stock, J.H. and Watson, M.W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179.
[87] Wachter, K.W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probab.* **6** 1–18.
[88] Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** 1342–1374.
[89] Wedin, P.Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111.
[90] Yu, Y., Wang, T. and Samworth, R.J. (2014). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323.
[91] Zhang, A., Cai, T.T. and Wu, Y. (2019). Heteroskedastic PCA: Algorithm, optimality, and applications. arXiv preprint, arXiv:1810.08316.
[92] Zhang, H., He, W., Zhang, L., Shen, H. and Yuan, Q. (2013). Hyperspectral image restoration using low-rank matrix recovery. *IEEE Trans. Geosci. Remote Sens.* **52** 4729–4743.

[93] Zhang, L., Zhou, W. and Wang, H. (2022). Supplement to "Non-Asymptotic Properties of Spectral Decomposition of Large Gram-Type Matrices and Applications." https://doi.org/10.3150/21-BEJ1384SUPP
[94] Zhang, X. and Cheng, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli* **24** 2640–2675.
[95] Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286.
[96] Zou, H. and Xue, L. (2018). A selective overview of sparse principal component analysis. *Proc. IEEE* **106** 1311–1320.

Received February 2020 and revised June 2021