数理解析研究所講究録
第1954巻 2015年 23-31

Reconstruction of a signal matrix for high-dimension, low-sample-size data

筑波大学・数理物質科学研究科　村山　航（Wataru Murayama）
Graduate School of Pure and Applied Sciences
University of Tsukuba

筑波大学・数学域　矢田　和善　(Kazuyoshi Yata)
Institute of Mathematics
University of Tsukuba

筑波大学・数学域　青嶋　誠　(Makoto Aoshima)
Institute of Mathematics
University of Tsukuba

Abstract: We consider recovering a signal matrix in high-dimension, low-sample-size (HDLSS) situations. We first consider using the conventional PCA to estimate a signal matrix and show that the induced estimator holds consistency properties under several conditions. We show that the estimator is directly affected by noise structures. In order to overcome the difficulty, we apply the noise-reduction (NR) methodology to a recovery of the signal matrix. We show that the NR method gives a preferable estimator of the signal matrix which holds the consistency properties under mild conditions. The NR method improves the accuracy of the conventional PCA successfully. Finally, we give several simulation results to recover a signal matrix.

Key words and phrases: HDLSS, Noise-reduction methodology, Large p small n, PCA.

1 Introduction

High-dimension, low-sample-size (HDLSS) data situations occur in many areas of modern science such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. The asymptotic studies of HDLSS data are becoming increasingly relevant. In recent years, substantial work has been done on HDLSS asymptotic theory. Hall et al. [7], Ahn et al. [1], and Yata and Aoshima [11] explored several types of geometric representations of HDLSS data. Jung and Marron [8] investigated the inconsistency of the eigenvalues and eigenvectors of the sample covariance matrix. Yata and Aoshima [11] gave consistent estimators for both the eigenvalues and eigenvectors together with the principal component (PC) scores by developing the *noise-reduction (NR) methodology*. The HDLSS asymptotic theory had been studied under the assumption that either the population distribution is Gaussian or the random variables in a sphered data matrix have a ρ-mixing dependency. However, Yata and Aoshima [9] provided asymptotic theory without assuming either the Gaussian assumption or the ρ-mixing condition. Moreover, Yata and Aoshima [10] created a new principal component analysis (PCA) called the *cross-data-matrix methodology* that is applicable to constructing an unbiased estimator in non-parametric settings. Aoshima and Yata [3, 4] developed a variety of high-dimensional statistical inference based on the geometric representations by using the cross-data-matrix methodology. See Aoshima and Yata [5, 6] for a review covering this field of research.

In this paper, we address the problem of recovering an unknown $d \times n$ low-rank matrix, $\boldsymbol{A} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_n]$. $\boldsymbol{A}$ is called the signal matrix. Let $r = \text{rank}(\boldsymbol{A})$. We assume r ($\le \min\{d, n\}$) is fixed. Suppose we have a $d \times n$ data matrix, $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$, where

$$\boldsymbol{X} = \sqrt{n}\boldsymbol{A} + \boldsymbol{W}. \tag{1}$$

Here, $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_n]$ is a $d \times n$ noise matrix, where $\boldsymbol{w}_j$, $j = 1, ..., n$, are independent and identically distributed (i.i.d.) as a d-dimensional distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_W$ $(\ge \boldsymbol{O})$. Note that $\boldsymbol{x}_j - \sqrt{n}\boldsymbol{a}_j$, $j = 1, ..., n$, are i.i.d. Let $\boldsymbol{\Sigma}_A = \boldsymbol{A}\boldsymbol{A}^T$. Then, it holds that $E(\boldsymbol{X}\boldsymbol{X}^T)/n = \boldsymbol{\Sigma}_A + \boldsymbol{\Sigma}_W$ $(= \boldsymbol{\Sigma}$, say). Andrey and Nobel [2] considered the model (1) in a high-dimensional setting, where the data dimension d and the sample size n increase at the same rate, i.e. $n/d \to c > 0$. They assumed that the elements of $\boldsymbol{W}$ are i.i.d. standard normal random variables. Note that the conditions such as "$n/d \to c > 0$" and the normality are quite strict in real high-dimensional analyses. In this paper, we consider the model (1) in HDLSS settings without assuming the severe conditions.

The eigen-decomposition of $\boldsymbol{\Sigma}_W$ is given by $\boldsymbol{\Sigma}_W = \boldsymbol{U}_W \boldsymbol{\Lambda}_W \boldsymbol{U}_W^T$, where $\boldsymbol{\Lambda}_W$ is a diagonal matrix of eigenvalues, $\lambda_{1(W)} \ge \cdots \ge \lambda_{d(W)} (\ge 0)$, and $\boldsymbol{U}_W$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{W} = \boldsymbol{U}_W \boldsymbol{\Lambda}_W^{1/2} \boldsymbol{Z}$. Then, $\boldsymbol{Z}$ is a $d \times n$ sphered data matrix from a distribution with the identity covariance matrix, $\boldsymbol{I}_n$. Here, we write $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_d]^T$ and $\boldsymbol{z}_j = (z_{j1}, ..., z_{jn})^T$, $j = 1, ..., d$. Note that $E(z_{jk} z_{j'k}) = 0$ $(j \ne j')$ and $\text{Var}(\boldsymbol{z}_j) = \boldsymbol{I}_n$. We assume that the fourth moments of each variable in $\boldsymbol{Z}$ are uniformly bounded. The singular value decomposition of $\boldsymbol{A}$ is given by $\boldsymbol{A} = \sum_{j=1}^r \lambda_{j(A)}^{1/2} \boldsymbol{u}_{j(A)} \boldsymbol{v}_{j(A)}^T$, where $\lambda_{1(A)}^{1/2} \ge \cdots \ge \lambda_{r(A)}^{1/2} (\ge 0)$ are singular values of $\boldsymbol{A}$ and $\boldsymbol{u}_{j(A)}$ (or $\boldsymbol{v}_{j(A)}$) denotes a unit left- (or right-) singular vector corresponding to $\lambda_{j(A)}^{1/2}$ $(j = 1, ..., r)$. In this paper, we assume the following model.

$$\limsup_{d\to\infty} \frac{\lambda_{1(A)}}{\lambda_{r(A)}} < \infty \quad \text{and} \quad \lim_{d\to\infty} \frac{\text{tr}(\boldsymbol{\Sigma}_W^2)}{\lambda_{r(A)}^2} = 0. \tag{2}$$

The model (2) is a special case of the power spiked model given by Yata and Aoshima [12]. Also, we assume that $\lambda_j(A)$s are distinct in the sense that

$$\liminf_{d\to\infty} |\lambda_{j(A)}/\lambda_{j'(A)} - 1| > 0$$

for all $j \ne j'$ $(\le r)$.

In Section 2, we consider using the conventional PCA to estimate $\boldsymbol{A}$ and show that the induced estimator holds consistency properties under several conditions. In Section 3, we apply the NR method instead and show that the NR method gives a preferable estimator which holds the consistency properties under mild conditions. The NR method improves the accuracy of the conventional PCA successfully. Finally, in Section 4, we give several simulation results to recover a signal matrix.

2 Reconstruction of A by conventional PCA

In this section, we consider recovering the signal matrix $\boldsymbol{A}$ by using the conventional PCA. The sample covariance matrix is given by $\boldsymbol{S} = n^{-1}\boldsymbol{X}\boldsymbol{X}^T$. We consider the dual sample covariance matrix defined by $\boldsymbol{S}_D = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$. Note that $\boldsymbol{S}_D$ and $\boldsymbol{S}$ share non-zero eigenvalues and rank($\boldsymbol{S}$) = rank($\boldsymbol{S}_D$) $\le \min\{d,n\}$. Let $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_{\min\{d,n\}} \ge 0$ be the eigenvalues of $\boldsymbol{S}_D$. The eigen-decompositions of $\boldsymbol{S}$ and $\boldsymbol{S}_D$ are given by $\boldsymbol{S} = \sum_{j=1}^{\min\{d,n\}} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$ and $\boldsymbol{S}_D = \sum_{j=1}^{\min\{d,n\}} \hat{\lambda}_j \hat{\boldsymbol{v}}_j \hat{\boldsymbol{v}}_j^T$ respectively, where $\hat{\boldsymbol{u}}_j$ (or $\hat{\boldsymbol{v}}_j$) denotes a unit left- (or right-) singular vector of $\boldsymbol{X}$ corresponding to $\hat{\lambda}_j$. Note that $\hat{\boldsymbol{u}}_j$ can be calculated by $\hat{\boldsymbol{u}}_j = (n\hat{\lambda}_j)^{-1/2}\boldsymbol{X}\hat{\boldsymbol{v}}_j$.

We reconstruct $\boldsymbol{A}$ by $\hat{\lambda}_j$s, $\hat{\boldsymbol{u}}_j$s and $\hat{\boldsymbol{v}}_j$s. We assume the following conditions as necessary:

(C-i) $$\frac{\sum_{s,t=1}^{d} \lambda_{s(W)}\lambda_{t(W)} E\{(z_{sk}^2-1)(z_{tk}^2-1)\}}{n\lambda_{r(A)}^2} = o(1);$$

(C-ii) $$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_W)}{n\lambda_{r(A)}} = o(1).$$

Let $\kappa_j = \mathrm{tr}(\boldsymbol{\Sigma}_W)/(n\lambda_{j(A)})$ for $j = 1,\ldots,r$. We have the following results.

Theorem 1. *For $j = 1,\ldots,r$, it holds that as $d \to \infty$ and $n \to \infty$*

$$\frac{\hat{\lambda}_j}{\lambda_{j(A)}} = 1 + \kappa_j + o_p(1), \quad \hat{\boldsymbol{u}}_j^T \boldsymbol{u}_{j(A)} = (1+\kappa_j)^{-1/2} + o_p(1) \quad \textit{and} \quad \hat{\boldsymbol{v}}_j^T \boldsymbol{v}_{j(A)} = 1 + o_p(1)$$

under (C-i).

Corollary 1. *For $j = 1,\ldots,r$, it holds that as $d \to \infty$ and $n \to \infty$*

$$\frac{\hat{\lambda}_j}{\lambda_{j(A)}} = 1 + o_p(1), \quad \hat{\boldsymbol{u}}_j^T \boldsymbol{u}_{j(A)} = 1 + o_p(1) \quad \textit{and} \quad \hat{\boldsymbol{v}}_j^T \boldsymbol{v}_{j(A)} = 1 + o_p(1)$$

under (C-i) and (C-ii).

Based on the theoretical background, we consider recovering the signal matrix $\boldsymbol{A}$ by $\hat{\boldsymbol{A}}_r = \sum_{j=1}^{r} \hat{\lambda}_j^{1/2} \hat{\boldsymbol{u}}_j \hat{\boldsymbol{v}}_j^T$. Then, we have the following results.

Theorem 2. *It holds that as $d \to \infty$ and $n \to \infty$*

$$||\hat{\boldsymbol{A}}_r - \boldsymbol{A}||_F^2 = r\frac{tr(\boldsymbol{\Sigma}_W)}{n} + o_p(\lambda_{r(A)})$$

under (C-i), where $||\cdot||_F$ denotes the Frobenius norm.

Corollary 2. *It holds that as $d \to \infty$ and $n \to \infty$*

$$||\hat{\boldsymbol{A}}_r - \boldsymbol{A}||_F^2 = o_p(\lambda_{r(A)})$$

under (C-i) and (C-ii).

It should be noted that (C-ii) is a necessary condition to claim the consistency property such as $||\hat{\boldsymbol{A}}_r - \boldsymbol{A}||_F^2/\lambda_{r(A)} = o_p(1)$.

3 Reconstruction of A by NR method

We consider applying the NR method by Yata and Aoshima [11] to recover the signal matrix $\boldsymbol{A}$. By using the NR method, we obtain an estimator of $\lambda_{i(A)}$ as

$$\acute{\lambda}_i = \hat{\lambda}_i - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{j=1}^{i} \hat{\lambda}_j}{n-i} \quad (i = 1, ..., n-1). \tag{3}$$

The following result claims that $\acute{\lambda}_i$ holds the consistency property without (C-ii). Remember that $\hat{\lambda}_i$ requires (C-ii) to hold the consistency property.

Theorem 3. *For $j = 1, ..., r$, it holds that as $d \to \infty$ and $n \to \infty$*

$$\frac{\acute{\lambda}_j}{\lambda_{j(A)}} = 1 + o_p(1)$$

under (C-i).

Now, we consider an adjustment of $\acute{\lambda}_i$s to estimate the signal matrix $\boldsymbol{A}$:

$$\acute{\lambda}_{i(r)} = \hat{\lambda}_i - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{j=1}^{r} \hat{\lambda}_j}{n-r} \quad (i = 1, ..., r). \tag{4}$$

We consider recovering $\boldsymbol{A}$ by $\acute{\boldsymbol{A}}_r = \sum_{j=1}^{r} \acute{\lambda}_{j(r)}^{1/2} \hat{\boldsymbol{u}}_j \hat{\boldsymbol{v}}_j^T$. Then, we have the following results.

Theorem 4. *It holds that as $d \to \infty$ and $n \to \infty$*

$$||\acute{\boldsymbol{A}}_r - \boldsymbol{A}||_F^2 = 2\sum_{i=1}^{r} \lambda_{i(A)}\Big(1 - \frac{1}{(1+\kappa_i)^{1/2}}\Big) + o_p(\lambda_{r(A)})$$

under (C-i).

Corollary 3. *It holds that as $d \to \infty$ and $n \to \infty$*

$$||\acute{\boldsymbol{A}}_r - \boldsymbol{A}||_F^2 = o_p(\lambda_{r(A)})$$

under (C-i) and (C-ii).

From Theorems 2 and 4, we compare $2\lambda_{i(A)}\{1 - 1/(1+\kappa_i)^{1/2}\}$ with $\lambda_{i(A)}\kappa_i$. It holds that $2\{1 - 1/(1+\kappa_i)^{1/2}\} < \kappa_i (i = 1, ..., r)$ for any $\kappa_i > 0$, so that $||\acute{\boldsymbol{A}}_r - \boldsymbol{A}||_F^2$ is smaller than $||\hat{\boldsymbol{A}}_r - \boldsymbol{A}||_F^2$ asymptotically. Thus, $\acute{\boldsymbol{A}}_r$ improves the error rate of $\hat{\boldsymbol{A}}_r$.

4 Simulations

We used computer simulations to compare the performance of $\acute{\boldsymbol{A}}_r$ with $\hat{\boldsymbol{A}}_r$. We set $r = 3$ and $\boldsymbol{\Sigma}_A = \mathrm{diag}(\lambda_{1(A)}, \lambda_{2(A)}, \lambda_{3(A)}, 0, ..., 0)$ with $\lambda_{1(A)} = 2d^{3/5}$, $\lambda_{2(A)} = 1.5d^{3/5}$and $\lambda_{3(A)} = d^{3/5}$. Note that the model (2) holds. We generated pseudo random vectors for $\boldsymbol{w}_j$, $j = 1, ..., n$, i.i.d. as a d-dimensional normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_W$ $(\geq \boldsymbol{O})$.

We considered two cases: (a) $\boldsymbol{\Sigma}_W = \boldsymbol{I}_d$ and (b) $\boldsymbol{\Sigma}_W = (\sigma_{ij})$ with $\sigma_{ij} = (0.3)^{|i-j|^{1/3}}$. Let $F(\boldsymbol{M}) = ||\boldsymbol{M} - \boldsymbol{A}||_F^2/\lambda_{r(A)}$ for any $d \times n$ matrix, $\boldsymbol{M}$. The findings were obtained by averaging the outcomes from 2000 ($= K$, say) replications. Under a fixed scenario, suppose that the k-th replication ends with estimates, $F(\hat{\boldsymbol{A}}_r)_k$ and $F(\acute{\boldsymbol{A}}_r)_k$, for $k = 1, ..., K$. Let us simply write $F_{\hat{A}} = K^{-1}\sum_{k=1}^K F(\hat{\boldsymbol{A}}_r)_k$ and $F_{\acute{A}} = K^{-1}\sum_{k=1}^K F(\acute{\boldsymbol{A}}_r)_k$. We also considered the Monte Carlo variability. Let $\mathrm{var}(F_{\hat{A}}) = (K-1)^{-1}\sum_{k=1}^K (F(\hat{\boldsymbol{A}}_r)_k - F_{\hat{A}})^2$ and $\mathrm{var}(F_{\acute{A}}) = (K-1)^{-1}\sum_{k=1}^K (F(\acute{\boldsymbol{A}}_r)_k - F_{\acute{A}})^2$. Fig. 1 shows the behaviors of $(F_{\hat{A}}, F_{\acute{A}})$ and $(\mathrm{var}(F_{\hat{A}}), \mathrm{var}(F_{\acute{A}}))$ for $(d, n) = (2^s, 3s)$, $s = 5, ..., 11$, in the case of (a). Fig. 2 shows them in the case of (b). The dashed lines denote the simulation results. In the left panels of each figure, we gave the corresponding theoretical values, $r\mathrm{tr}(\boldsymbol{\Sigma}_W)/(n\lambda_r)$ and $2\sum_{i=1}^r \lambda_{i(A)}\{1 - (1+\kappa_i)^{-1/2}\}/\lambda_r$, that are denoted by the solid lines. See Theorems 2 and 4 for the details. The simulation results appeared close to the theoretical values and it seemed to be good approximations. As expected theoretically, we observed that the estimates by the NR method give more preferable performances both for (a) and (b) compared to the conventional PCA.

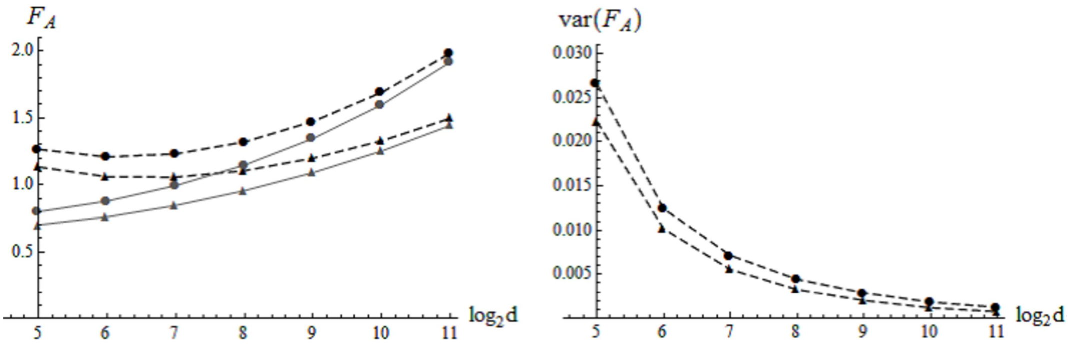

Figure 1: The behaviors of two estimates, $\hat{\boldsymbol{A}}_r$ denoted by ● and $\acute{\boldsymbol{A}}_r$ denoted by ▲, when $\boldsymbol{\Sigma}_W = \boldsymbol{I}_d$. The values of $F_{\hat{A}}$ and $F_{\acute{A}}$ are denoted by the dashed lines in the left panel. The values of their sample variances, $\mathrm{var}(F_{\hat{A}})$ and $\mathrm{var}(F_{\acute{A}})$, are denoted by the dashed lines in the right panel. The theoretical values, $r\mathrm{tr}(\boldsymbol{\Sigma}_W)/(n\lambda_r)$ and $2\sum_{i=1}^r \lambda_{i(A)}\{1 - (1+\kappa_i)^{-1/2}\}/\lambda_r$, are denoted by the solid lines in the left panel.

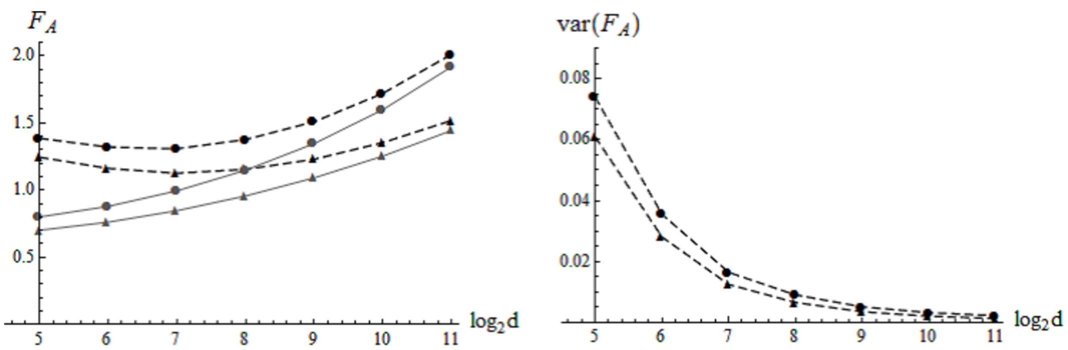

Figure 2: (Continued). When $\boldsymbol{\Sigma}_W = (\sigma_{ij})$ with $\sigma_{ij} = (0.3)^{|i-j|^{1/3}}$.

A Appendix

Throughout, let $\boldsymbol{e}_n = (e_1, ..., e_n)^T$ be an arbitrary unit random vector.

Lemma 1. *It holds that as $d \to \infty$ and $n \to \infty$*

$$\boldsymbol{e}_n^T \frac{\boldsymbol{W}^T\boldsymbol{W}}{n\lambda_{j(A)}}\boldsymbol{e}_n = \frac{tr(\boldsymbol{\Sigma}_W)}{n\lambda_{j(A)}} + o_p(1)$$

for j $(\le r)$ under (C-i).

Proof. We write that

$$\boldsymbol{e}_n^T \frac{\boldsymbol{W}^T\boldsymbol{W}}{n\lambda_{j(A)}}\boldsymbol{e}_n = \boldsymbol{e}_n^T(n^{-1}\sum_{s=1}^{d}\lambda_{s(W)}\boldsymbol{z}_s\boldsymbol{z}_s^T)\boldsymbol{e}_n = \boldsymbol{e}_n^T\{n^{-1}\sum_{s=1}^{d}\lambda_{s(W)}(\boldsymbol{z}_s\boldsymbol{z}_s^T - \boldsymbol{I}_n)\}\boldsymbol{e}_n + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_W)}{n}.$$

From Lemma 5 given in Yata and Aoshima [12], it holds that as $d \to \infty$ and $n \to \infty$

$$\boldsymbol{e}_n^T \frac{n^{-1}\sum_{s=1}^{d}\lambda_{s(W)}(\boldsymbol{z}_s\boldsymbol{z}_s^T - \boldsymbol{I}_n)}{\lambda_j}\boldsymbol{e}_n = o_p(1)$$

for j $(\le r)$ under (C-i). Thus it concludes the result. □

Lemma 2. *It holds that as $d \to \infty$ and $n \to \infty$*

$$\frac{\boldsymbol{u}_{i(A)}^T\boldsymbol{W}\boldsymbol{e}_n}{n^{1/2}} = o_p(\lambda_{r(A)}^{1/2}),\ i = 1, ..., r.$$

Proof. We write that $\boldsymbol{u}_{i(A)}^T\boldsymbol{W}\boldsymbol{e}_n = \sum_{k=1}^{n} e_k\boldsymbol{w}_k^T\boldsymbol{u}_{i(A)}$. Note that $\lambda_{1(W)} = o(\lambda_{r(A)})$ as $d \to \infty$ from (2). By using Markov's inequality, for any $\tau > 0$ and $i = 1, ..., r$, we have that as $d \to \infty$ and $n \to \infty$

$$P\Big(\sum_{k=1}^{n}(\boldsymbol{w}_k^T\boldsymbol{u}_{i(A)})^2/n \ge \tau\lambda_{r(A)}\Big) \le \frac{E\{\sum_{k=1}^{n}(\boldsymbol{w}_k^T\boldsymbol{u}_{i(A)})^2\}}{\tau n\lambda_{r(A)}} = \frac{\boldsymbol{u}_{i(A)}^T\boldsymbol{\Sigma}_W\boldsymbol{u}_{i(A)}}{\tau\lambda_{r(A)}} \le \frac{\lambda_{1(W)}}{\tau\lambda_{r(A)}} = o(1)$$

from the fact that $\boldsymbol{u}_{i(A)}^T\boldsymbol{\Sigma}_W\boldsymbol{u}_{i(A)} \le \lambda_{1(W)}$. Then, by noting that

$$\Big|\sum_{k=1}^{n} e_k(\boldsymbol{w}_k^T\boldsymbol{u}_{i(A)})/n^{1/2}\Big| \le \Big\{\sum_{k=1}^{n} e_k^2\Big\}^{1/2}\Big\{\sum_{k=1}^{n}(\boldsymbol{w}_k^T\boldsymbol{u}_{i(A)})^2/n\Big\}^{1/2} = \Big\{\sum_{k=1}^{n}(\boldsymbol{w}_k^T\boldsymbol{u}_{i(A)})^2/n\Big\}^{1/2},$$

we can conclude the result. □

Proof of Theorem 1. We write that for $j = 1, ..., r$

$$\frac{\hat{\lambda}_j}{\lambda_{j(A)}} = \hat{\boldsymbol{v}}_j^T \frac{\boldsymbol{S}_D}{\lambda_{j(A)}} \hat{\boldsymbol{v}}_j = \hat{\boldsymbol{v}}_j^T \frac{(\boldsymbol{A} + \boldsymbol{W}/n^{1/2})^T (\boldsymbol{A} + \boldsymbol{W}/n^{1/2})}{\lambda_{j(A)}} \hat{\boldsymbol{v}}_j$$
$$= \hat{\boldsymbol{v}}_j^T \frac{\boldsymbol{A}^T\boldsymbol{A}}{\lambda_{j(A)}} \hat{\boldsymbol{v}}_j + 2\hat{\boldsymbol{v}}_j^T \frac{\boldsymbol{A}^T\boldsymbol{W}}{n^{1/2}\lambda_{j(A)}} \hat{\boldsymbol{v}}_j + \hat{\boldsymbol{v}}_j^T \frac{\boldsymbol{W}^T\boldsymbol{W}}{n\lambda_{j(A)}} \hat{\boldsymbol{v}}_j. \quad (5)$$

We note that $|\hat{\boldsymbol{v}}_j^T \boldsymbol{A}^T \boldsymbol{W} \hat{\boldsymbol{v}}_j| \le \sum_{i=1}^r \lambda_{i(A)}^{1/2} |\boldsymbol{u}_{i(A)}^T \boldsymbol{W} \hat{\boldsymbol{v}}_j|$ for $j = 1, ..., r$. From Lemma 2 and (2), it holds that as $d \to \infty$ and $n \to \infty$

$$\hat{\boldsymbol{v}}_j^T \boldsymbol{A}^T \boldsymbol{W} \hat{\boldsymbol{v}}_j / n^{1/2} = o_p(\lambda_{r(A)}^{1/2}) \quad (6)$$

for $j = 1, ..., r$. Then, by combining (5) with Lemma 1 and (6), it holds that for $j = 1, ..., r$

$$\frac{\hat{\lambda}_j}{\lambda_{j(A)}} = \hat{\boldsymbol{v}}_j^T \frac{\sum_{s=1}^r \lambda_{s(A)} \boldsymbol{v}_{s(A)} \boldsymbol{v}_{s(A)}^T}{\lambda_{j(A)}} \hat{\boldsymbol{v}}_j + \frac{\text{tr}(\boldsymbol{\Sigma}_W)}{n\lambda_{j(A)}} + o_p(1) = 1 + \frac{\text{tr}(\boldsymbol{\Sigma}_W)}{n\lambda_{j(A)}} + o_p(1)$$

under (C-i). Thus, we have that $\hat{\boldsymbol{v}}_j^T \boldsymbol{v}_{j(A)} = 1 + o_p(1)$ for $j = 1, ..., r$. For $\hat{\boldsymbol{u}}_j$s, from Lemma 2, under (C-i), it holds that as $d \to \infty$ and $n \to \infty$

$$\boldsymbol{u}_{j(A)}^T \hat{\boldsymbol{u}}_j = \hat{\lambda}_j^{-1/2} \lambda_{j(A)}^{1/2} \boldsymbol{v}_{j(A)}^T \hat{\boldsymbol{v}}_j + (n\hat{\lambda}_j)^{-1/2} \boldsymbol{u}_{j(A)}^T \boldsymbol{W} \hat{\boldsymbol{v}}_j = \{1 + \text{tr}(\boldsymbol{\Sigma}_W)/(n\lambda_{j(A)})\}^{-1/2} + o_p(1)$$

for $j = 1, ..., r$. Thus it concludes the results. □

Proof of Corollary 1. Note that $\text{tr}(\boldsymbol{\Sigma}_W)/n = o(\lambda_{r(A)})$ under (C-ii). From Theorem 1, we can conclude the result. □

Proofs of Theorem 2 and Corollary 2. From Theorem 1, under (C-i), we have that

$$||\hat{\boldsymbol{A}}_r - \boldsymbol{A}||_F^2 = ||\sum_{s=1}^r (\hat{\lambda}_s^{1/2} \hat{\boldsymbol{u}}_s \hat{\boldsymbol{v}}_s^T - \lambda_{s(A)}^{1/2} \boldsymbol{u}_{s(A)} \boldsymbol{v}_{s(A)}^T)||_F^2 = r\frac{\text{tr}(\boldsymbol{\Sigma}_W)}{n} + o_p(\lambda_{r(A)}).$$

It concludes the result. □

Lemma 3. *It holds that as $d \to \infty$ and $n \to \infty$*

$$\frac{tr(\boldsymbol{S}_D) - \sum_{i=1}^j \hat{\lambda}_i}{n-j} = \frac{tr(\boldsymbol{\Sigma}_W)}{n} + o_p(\lambda_{r(A)}) \quad \text{for } j = 1, ..., r$$

under (C-i).

Proof. We write that $\text{tr}(\boldsymbol{W}^T\boldsymbol{W}/n) - \text{tr}(\boldsymbol{\Sigma}_W) = \sum_{s=1}^d \lambda_{s(W)} \sum_{k=1}^n (z_{sk}^2 - 1)/n$. Then, it holds that as $d \to \infty$ and $n \to \infty$

$$E[\{\text{tr}(\boldsymbol{W}^T\boldsymbol{W}/n) - \text{tr}(\boldsymbol{\Sigma}_W)\}^2] = \sum_{r,s=1}^d \lambda_{r(W)} \lambda_{s(W)} E\{(z_{sk}^2 - 1)(z_{sk}^2 - 1)\}/n = o(\lambda_{r(A)}^2)$$

under (C-i). Hence, by using Chebyshev's inequality, for any $\tau > 0$, we have that as $d \to \infty$ and $n \to \infty$

$$P\Big(|\mathrm{tr}(\boldsymbol{W}^T\boldsymbol{W}/n) - \mathrm{tr}(\boldsymbol{\Sigma}_W)| \geq \tau\lambda_{r(A)}\Big) \leq \frac{E[\{\mathrm{tr}(\boldsymbol{W}^T\boldsymbol{W}/n) - \mathrm{tr}(\boldsymbol{\Sigma}_W)\}^2]}{\tau^2\lambda_{r(A)}^2} = o(1)$$

under (C-i). Thus it follows that $\mathrm{tr}(\boldsymbol{W}^T\boldsymbol{W}/n) = \mathrm{tr}(\boldsymbol{\Sigma}_W) + o_p(\lambda_{r(A)})$. On the other hand, from (2), we have that

$$\begin{aligned} E\{\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{W})^2\} &= E\Big\{\Big(\sum_{i=1}^n \boldsymbol{a}_i^T\boldsymbol{w}_i\Big)^2\Big\} = \sum_{i=1}^n \boldsymbol{a}_i^T\boldsymbol{\Sigma}_W\boldsymbol{a}_i = \mathrm{tr}(\boldsymbol{\Sigma}_W\boldsymbol{\Sigma}_A) \\ &\leq \sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_W^2)\mathrm{tr}(\boldsymbol{\Sigma}_A^2)} = O(\mathrm{tr}(\boldsymbol{\Sigma}_W^2)^{1/2}r\lambda_{1(A)}) = o(\lambda_{r(A)}^2), \end{aligned}$$

so that $\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{W}) = o_p(\lambda_{r(A)})$. Then, we have that

$$\begin{aligned} \mathrm{tr}(\boldsymbol{S}_D) &= \mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A}) + 2\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{W})/n^{1/2} + \mathrm{tr}(\boldsymbol{W}^T\boldsymbol{W})/n \\ &= \sum_{i=1}^r \lambda_{i(A)} + \mathrm{tr}(\boldsymbol{\Sigma}_W) + o_p(\lambda_{r(A)}). \end{aligned} \tag{7}$$

under (C-i). From Theorem 1 and (7), it holds that for j $(\leq r)$

$$\frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{i=1}^j \hat{\lambda}_i}{n-j} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_W)}{n} + o_p(\lambda_{r(A)})$$

under (C-i). It concludes the result. □

Proof of Theorem 3. By combining Theorem 1 with Lemma 3, we can conclude the result. □

Proofs of Theorem 4 and Corollary 3. By combining Theorems 1 and 3, we can conclude the results. □

Acknowledgment

Research of the first author was partially supported by Grant-in-Aid for Young Scientists (B), Japan Society for the Promotion of Science (JSPS), under Contract Number 26800078. Research of the second author was partially supported by Grants-in-Aid for Scientific Research (B) and Challenging Exploratory Research, JSPS, under Contract Numbers 22300094 and 26540010.

References

[1] J. Ahn, J.S. Marron, K.M. Muller, Y.-Y. Chi, The high-dimension, low-sample-size geometric representation holds under mild conditions, Biometrika 94 (2007) 760-766.

[2] A. Andrey, A. Nobel, Reconstruction of a low-rank matrix in the presence of Gaussian noise, J. Multivariate Anal. 118 (2013) 67-76.

[3] M. Aoshima, K. Yata, Two-stage procedures for high-dimensional data, Sequential Anal. (*Editor's special invited paper*) 30 (2011) 356-399.

[4] M. Aoshima, K. Yata, Authors' response, Sequential Anal. 30 (2011) 432-440.

[5] M. Aoshima, K. Yata, Invited review article: Statistical inference in high-dimension, low-sample-size settings, Sugaku 65 (2013) 225-247.

[6] M. Aoshima, K. Yata, The JSS research prize lecture: Effective methodologies for high-dimensional data, J. Japan Statist. Soc. Ser. J 43 (2013) 123-150.

[7] P. Hall, J.S. Marron, A. Neeman, Geometric representation of high dimension, low sample size data, J. R. Statist. Soc. Ser. B 67 (2005) 427-444.

[8] S. Jung, J.S. Marron, PCA consistency in high dimension, low sample size context, Ann. Statist. 37 (2009) 4104-4130.

[9] K. Yata, M. Aoshima, PCA consistency for non-Gaussian data in high dimension, low sample size context, Commun. Statist. Theory Methods, Special Issue Honoring S. Zacks (Ed. N. Mukhopadhyay) 38 (2009) 2634-2652.

[10] K. Yata, M. Aoshima, Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, J. Multivariate Anal. 101 (2010) 2060-2077.

[11] K. Yata, M. Aoshima, Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, J. Multivariate Anal. 105 (2012) 193-215.

[12] K. Yata, M. Aoshima, PCA consistency for the power spiked model in high-dimensional settings, J. Multivariate Anal. 122 (2013) 334-354.