High-dimensional covariance matrix estimation under the SSE model

筑波大学大学院・数理物質科学研究科　　小西　啓介（Keisuke Konishi）
Graduate School of Pure and Applied Sciences
University of Tsukuba
筑波大学・数理物質系　　矢田　和善（Kazuyoshi Yata）
Institute of Mathematics
University of Tsukuba
筑波大学・数理物質系　　青嶋　誠（Makoto Aoshima）
Institute of Mathematics
University of Tsukuba

Abstract

In this paper, we consider the estimation for the inverse matrix of a high-dimensional covariance matrix under the strongly spiked eigenvalue model. One of the well-known estimation methods is the principal orthogonal complement thresholding (POET) given by Fan et al. [5]. We show that the POET has consistency properties only under several severe conditions in high-dimensional settings. In order to overcome the difficulty, we consider applying the noise-reduction (NR) method given by Yata and Aoshima [8, 9] to the POET. We propose a new estimation of the inverse covariance matrix called the NR-POET. We compare the performance of the NR-POET with the POET by several simulations.

1 Introduction

One of the features of high-dimensional data is that the data dimension d is high, however, the sample size n is low. This is the so-called "HDLSS" or "large d, small n" data. Such data situations appear in many fields of modern science such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. For HDLSS data, the sample covariance matrix does not have its inverse matrix. The estimation for the inverse matrix of a covariance matrix is a crucial issue for high-dimensional data analyses, especially for pathway analysis and graphical modeling.

Bickel and Levina [4] gave a thresholding estimator for the inverse matrix of a covariance matrix when the covariance matrix is sparse and its eigenvalues are bounded. However, such sparsity conditions are severe for actual data and often out of touch with reality. In fact, Aoshima and Yata [1, 2, 3] and Yata and Aoshima [8, 9] showed that the first several eigenvalues diverge as d grows and the bounded-eigenvalues condition is quite strict for microarray data sets. Fan et al. [5] proposed a different thresholding estimator called the principal orthogonal complement thresholding (POET) under the assumption that the first

several eigenvalues diverge rapidly at the rate of d. Unfortunately, the assumption required in the POET cannot express the structure of actual data.

Suppose we have a $d \times n$ data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, where $\boldsymbol{x}_i,\ i = 1, \ldots, n\ (< d)$, are independent and identically distributed (i.i.d.) as a d-dimensional distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. We denote the eigen-decomposition of $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ and $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_d]$ is an orthogonal matrix with eigenvectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_d$ corresponding to the $\lambda_1, \ldots, \lambda_d$. Let σ_{ij} be the (i, j) element of $\boldsymbol{\Sigma}$ for $i, j = 1, \ldots, d$. We assume that $\sigma_{jj} \in (0, \infty)$ as $d \to \infty$ for all j. For a function, $f(\cdot)$, "$f(d) \in (0, \infty)$ as $d \to \infty$" implies that $\liminf_{d\to\infty} f(d) > 0$ and $\limsup_{d\to\infty} f(d) < \infty$.

Let $\boldsymbol{x}_j = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z}_j$, where $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{dj})^T$ is considered as a sphered data vector having the zero mean vector and identity covariance matrix. We assume that

(C-i) $\limsup_{d\to\infty} E(z_{ri}^4) < \infty$ for all r, and
$E(z_{ri}^2 z_{si}^2) = E(z_{ri}^2)E(z_{si}^2) = 1$ and $E(z_{ri}z_{si}z_{ti}z_{ui}) = 0$ for all $r \neq s, t, u$.

When $\boldsymbol{x}_j$s are Gaussian, (C-i) naturally holds.

The sample covariance matrix is given by $\boldsymbol{S} = n^{-1}\boldsymbol{X}\boldsymbol{X}^T$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d \geq 0$ be the eigenvalues of $\boldsymbol{S}$. Then, we denote the eigen-decomposition of $\boldsymbol{S}$ by

$$\boldsymbol{S} = \sum_{i=1}^{d} \hat{\lambda}_i \hat{\boldsymbol{h}}_i \hat{\boldsymbol{h}}_i^T,$$

where $\hat{\boldsymbol{h}}_i$ is a unit eigenvector corresponding to the $\hat{\lambda}_i$. The dual sample covariance matrix is given by $\boldsymbol{S}_D = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$. We have the eigen-decomposition of $\boldsymbol{S}_D$ by

$$\boldsymbol{S}_D = \sum_{i=1}^{n} \hat{\lambda}_i \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i^T,$$

where $\hat{\boldsymbol{u}}_i$ is a unit eigenvector corresponding to the $\hat{\lambda}_i$.

We assume the following spiked model for the eigenvalues of $\boldsymbol{\Sigma}$:

(C-ii) $\dfrac{\lambda_i}{d^{1/2}} \to \infty$ as $d \to \infty$ for $i = 1, \ldots, m$, and $\lambda_i \in (0, \infty)$ as $d \to \infty$ for all $i \geq m + 1$.
Here, m is a positive and fixed integer. When $m \geq 2$, $\lambda_1, ..., \lambda_m$ are distinct in the sense that

$$\liminf_{d\to\infty}(\lambda_i/\lambda_j - 1) > 0 \quad \text{for } 1 \leq i < j \leq m.$$

Note that (C-ii) is one of the strongly spiked eigenvalue models given by Aoshima and Yata [1]. See Remark 1.2. We devide $\boldsymbol{\Sigma}$ into $\boldsymbol{\Sigma}_1 = \sum_{j=1}^{m} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T$ and $\boldsymbol{\Sigma}_2 = \sum_{j=m+1}^{d} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T$, so that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$. Here, $\boldsymbol{\Sigma}_1$ is regarded as the signal and $\boldsymbol{\Sigma}_2$ is regarded as the noise.

Remark 1.1. When we consider a spiked model such as

$$\lambda_i = a_i d^{\alpha_i} \ (i = 1, \ldots, m) \quad \text{and} \quad \lambda_i = c_i \ (i = m+1, \ldots, d) \tag{1}$$

with positive and fixed constants, a_is, c_is and α_is. Note that (C-ii) is met when $\alpha_m > 0.5$. For instance, when we analyze a microarray data set, we find several gene networks and each network consists of genes that are highly correlated to each other. The high correlation is one of the reasons why strongly spiked eigenvalues appear in high-dimensional data analyses.

Remark 1.2. Aoshima and Yata [1] showed that the asymptotic normality of high-dimensional statistics cannot be established under the following model called the "strongly spiked eigenvalue (SSE) model":

$$\liminf_{d\to\infty} \left\{ \frac{\lambda_1}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^{1/2}} \right\} > 0.$$

They gave a data transformation technique from the SSE model to the non-SSE model.

This paper is organized as follows: In Section 2, we consider the POET to construct an estimator of $\boldsymbol{\Sigma}^{-1}$ and show that the POET has consistency properties under several severe conditions. In Section 3, we introduce the noise-reduction (NR) method that was given by Yata and Aoshima [8, 9]. The NR method is a new PCA having consistency properties for high-dimensional data. In Section 4, we consider applying the NR method to the POET for the inverse matrix estimation. We propose a new estimation of the inverse covariance matrix, called the NR-POET. Finally, in Section 5, we compare the performance of the NR-POET with POET by several simulations.

2 POET and its asymptotic properties

In this section, we introduce the principal orthogonal complement thresholding (POET) given by Fan et al. [5]. Let σ_{2ij} be the (i,j) element of $\boldsymbol{\Sigma}_2$. Let $\tau_{d,h} = \max_{1\le i\le d} \sum_{j=1}^{d} |\sigma_{2ij}|^h$ for $0 \le h \le 1$. Here, $\tau_{d,h}$ is the sparsity measure given by Bickel and Levina [4]. If $\tau_{d,h}$ is much smaller than d for a constant $h \in [0,1)$, $\boldsymbol{\Sigma}_2$ is considered as sparse in the sense that many elements of $\boldsymbol{\Sigma}_2$ are very small. We assume $\limsup_{d\to\infty} \tau_{d,h} < \infty$ for a constant $h \in [0,1)$. Let $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ be the smallest and largest eigenvalues of any positive definite matrix, $\boldsymbol{M}$. Note that $\lambda_{\max}(\boldsymbol{\Sigma}_2) = O(\tau_{d,h})$ and $\limsup_{d\to\infty} \lambda_{\max}(\boldsymbol{\Sigma}_2) < \infty$. We assume that $\liminf_{d\to\infty} \lambda_{\min}(\boldsymbol{\Sigma}_2) > 0$.

Let

$$\boldsymbol{w}_l = (w_{1l}, \ldots, w_{dl})^T = \left(\boldsymbol{I}_d - \sum_{j=1}^{m} \boldsymbol{h}_j \boldsymbol{h}_j^T \right) \boldsymbol{x}_l \quad \text{for} \quad l = 1, \ldots, n$$

and

$$\hat{\boldsymbol{w}}_l = (\hat{w}_{1l}, \ldots, \hat{w}_{dl})^T = \left(\boldsymbol{I}_d - \sum_{j=1}^{m} \hat{\boldsymbol{h}}_j \hat{\boldsymbol{h}}_j^T \right) \boldsymbol{x}_l \quad \text{for} \quad l = 1, \ldots, n.$$

Let $I(\cdot)$ be the indicator function. A thresholding operator is defined by

$$T(\boldsymbol{M}) = [m_{ij}\{I(i=j) + I(i \neq j)I(|m_{ij}| \geq t_{ij})\}]$$

for any symmetric matrix $\boldsymbol{M} = (m_{ij})$ and $t_{ij} > 0,\ i \neq j$. Fan et al. [5] considered estimating $\boldsymbol{\Sigma}_2$ by $T\big(\widehat{\boldsymbol{\Sigma}}_2\big)$ with $\widehat{\boldsymbol{\Sigma}}_2 = \sum_{l=1}^{n} \hat{\boldsymbol{w}}_l \hat{\boldsymbol{w}}_l^T/n$ and

$$t_{ij} = C\sqrt{\hat{\theta}_{ij}}\Big(d^{-1/2} + \sqrt{n^{-1}\log d}\Big) \text{ for all } i \neq j. \tag{2}$$

Here, $C(>0)$ is a sufficeintly large constant and $\hat{\theta}_{ij} = n^{-1}\sum_{s=1}^{n}(\hat{w}_{is}\hat{w}_{js} - n^{-1}\sum_{t=1}^{n}\hat{w}_{it}\hat{w}_{jt})^2$. Then, they gave an estimator of $\boldsymbol{\Sigma}$ by

$$\widehat{\boldsymbol{\Sigma}} = \sum_{j=1}^{m} \hat{\lambda}_j \hat{\boldsymbol{h}}_j \hat{\boldsymbol{h}}_j^T + T\big(\widehat{\boldsymbol{\Sigma}}_2\big). \tag{3}$$

They assumed the following spiked model:

(C-ii') $\dfrac{\lambda_i}{d} \in (0,\infty)$ as $d \to \infty$ for $i = 1,\ldots,m$, and $\lambda_i \in (0,\infty)$ as $d \to \infty$ for all $i \geq m+1$.

Note that (C-ii') is met when $\alpha_m = 1$ in (1). Thus (C-ii') is much stricter than (C-ii). Let $v = \min\{d, n\}$. We denote the frobenius and spectral norms by $||\cdot||_F$ and $||\cdot||$, that is, $||\boldsymbol{M}||_F = \{\mathrm{tr}(\boldsymbol{M}^T\boldsymbol{M})\}^{1/2}$ and $||\boldsymbol{M}|| = \{\lambda_{\max}(\boldsymbol{M}^T\boldsymbol{M})\}^{1/2}$ for any $d \times d$ matrix, $\boldsymbol{M}$. Then, the following result was given by Fan et al. [5].

Theorem 2.1 ([5]). *Assume (C-i) and (C-ii'). Then, under some regularity conditions, for a sufficiently large constant $C(>0)$ it holds that as $v \to \infty$*

$$||\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}_d||_F/\sqrt{d} = o_P(1) \quad \textit{and} \quad ||\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}|| = o_P(1).$$

It should be noted that (C-ii') is quite strict for high-dimensional data.

3 NR method and its asymptotic properties

In this section, we introduce the noise-reduction (NR) method given by Yata and Aoshima [8, 9]. Let $\delta_j = \lambda_j^{-1}\sum_{s=m+1}^{d}\lambda_s/n$ for $j = 1,\ldots,m$. Aoshima and Yata [1] and Yata and Aoshima [9] gave the following result when $v \to \infty$.

Proposition 3.1 ([1, 9]). *Assume (C-i) and (C-ii). It holds for $j = 1, ..., m$, that $\hat{\lambda}_j/\lambda_j = 1 + \delta_j + O_P(n^{-1/2})$ and $(\hat{\boldsymbol{h}}_j^T\boldsymbol{h}_j)^2 = (1+\delta_j)^{-1} + O_P(n^{-1/2})$ as $v \to \infty$.*

If $\delta_j \to \infty$ as $v \to \infty$, $\hat{\lambda}_j$ and $\hat{\boldsymbol{h}}_j$ are strongly inconsistent in the sense that $\lambda_j/\hat{\lambda}_j = o_P(1)$ and $(\hat{\boldsymbol{h}}_j^T\boldsymbol{h}_j)^2 = o_P(1)$. See Jung and Marron [6] for the concept of the strong inconsistency. In order to overcome the curse of dimensionality, Yata and Aoshima [8, 9] proposed an eigenvalue

estimation called the noise-reduction (NR) method, which was brought about by a geometric representation of $\boldsymbol{S}_D$. If one applies the NR method, the λ_j is estimated by

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{l=1}^{j} \hat{\lambda}_l}{n-j} \quad (j = 1, ..., n-1). \tag{4}$$

Note that $\tilde{\lambda}_j \geq 0$ w.p.1 for $j = 1, ..., n-1$, and the second term in (4) is an estimator of $\lambda_j \delta_j$. When applying the NR method to the PC direction vector, one obtains

$$\tilde{\boldsymbol{h}}_j = (n\tilde{\lambda}_j)^{-1/2} \boldsymbol{X} \hat{\boldsymbol{u}}_j$$

for $j = 1, ..., n-1$. Then, we have the following result.

Proposition 3.2 ([1, 9]). *Assume (C-i) and (C-ii). It holds for $j = 1, ..., k$, that $\tilde{\lambda}_j / \lambda_j = 1 + O_P(n^{-1/2})$ and $(\tilde{\boldsymbol{h}}_j^T \boldsymbol{h}_j)^2 = 1 + O_P(n^{-1})$ as $v \to \infty$.*

Thus, $\tilde{\lambda}_j$ and $\tilde{\boldsymbol{h}}_j$ have the consistency properties even when $\delta_j \to \infty$ and (C-ii) is met.

Remark 3.1. Wang and Fan [7] proposed the following estimator of $\boldsymbol{\Sigma}$ by using the NR method:

$$\widehat{\boldsymbol{\Sigma}} = \sum_{j=1}^{m} \tilde{\lambda}_j \hat{\boldsymbol{h}}_j \hat{\boldsymbol{h}}_j^T + T\big(\widehat{\boldsymbol{\Sigma}}_2\big). \tag{5}$$

They called this estimation method the Shrinkage-POET (S-POET).

For estimating $\boldsymbol{x}_l^T \boldsymbol{h}_j$, Aoshima and Yata [1] showed that $\boldsymbol{x}_l^T \hat{\boldsymbol{h}}_j$ and even $\boldsymbol{x}_l^T \tilde{\boldsymbol{h}}_j$ involve a huge bias. In order to overcome the inconvenience, Aoshima and Yata [1] gave the modofied NR method. According to [1], we modify $\boldsymbol{x}_l^T \tilde{\boldsymbol{h}}_j$ as $\boldsymbol{x}_l^T \tilde{\boldsymbol{h}}_{jl}$ by

$$\tilde{\boldsymbol{h}}_{jl} = \Big(\frac{n}{n-1}\Big) \frac{\boldsymbol{X} \hat{\boldsymbol{u}}_{jl}}{(n\tilde{\lambda}_j)^{1/2}} = \frac{n^{1/2} \boldsymbol{X} \hat{\boldsymbol{u}}_{jl}}{(n-1)\tilde{\lambda}_j^{1/2}},$$

where

$$\hat{\boldsymbol{u}}_{jl} = (\hat{u}_{j1}, ..., \hat{u}_{jl-1}, 0, \hat{u}_{jl+1}, ..., \hat{u}_{jn})^T.$$

Note that $\sum_{l=1}^{n} \hat{\boldsymbol{u}}_{jl}/n = \{(n-1)/n\} \hat{\boldsymbol{u}}_j$. We give the following R-code to calculate $\boldsymbol{x}_l^T \tilde{\boldsymbol{h}}_{jl}$:

```
xTh <- function(X, m, MeanZero=F){
    d <- dim(X)[1]
    n <- dim(X)[2]
    if (MeanZero){
        r <- min(n-1, d)
        Sd <- t(X) %*% X / n
        eig <- eigen(Sd)
        dualval <- eig$values[1:m]
        dualvec <- eig$vectors
        ans <- matrix(0, n, m)
        c <- sqrt(n) / (n-1)
        for (i in 1:m){
            nrmval <- dualval[i] - (sum(diag(Sd)) - sum(dualval[1:i])) / (n-i)
            u_hat <- dualvec[, i]
            for (j in 1:n){
                u_hat[j] <- 0
                nrmvec_self <- c * X %*% u_hat / sqrt(nrmval)
                ans[j, i] <- as.numeric(t(nrmvec_self) %*% X[, j])
                u_hat <- dualvec[, i]
            }
        }
    } else {
        r <- min(n-2, d)
        X <- sweep(X, 1, apply(X, 1, mean), '-')
        Sd <- t(X) %*% X / (n-1)
        eig <- eigen(Sd)
        dualval <- eig$values[1:m]
        dualvec <- eig$vectors
        ans <- matrix(0, n, m)
        c <- sqrt(n-1) / (n-2)
        for (i in 1:m){
            nrmval <- dualval[i] - (sum(diag(Sd)) - sum(dualval[1:i])) / (n-i
                -1)
            u_hat <- dualvec[, i]
            for (j in 1:n){
                u_hat[j] <- 0
                nrmvec_self <- c * X %*% u_hat / sqrt(nrmval)
                ans[j, i] <- as.numeric(t(nrmvec_self) %*% X[, j])
                u_hat <- dualvec[, i]
            }
        }
    }
    return(ans)
}
```

4 NR-POET

In this section, we propose a new estimation of $\mathbf{\Sigma}^{-1}$ by applying the NR method to the POET. By using the modified NR method in Section 3, we estimate $\boldsymbol{w}_l$ by

$$\tilde{\boldsymbol{w}}_l = (\tilde{w}_{1l}, \ldots, \tilde{w}_{dl})^T = \left(\boldsymbol{I}_d - \sum_{j=1}^{m} \hat{\boldsymbol{h}}_j \tilde{\boldsymbol{h}}_{jl}^T \right) \boldsymbol{x}_l.$$

We consider estimating $\mathbf{\Sigma}_2$ by $T(\widetilde{\mathbf{\Sigma}}_2)$, where $\widetilde{\mathbf{\Sigma}}_2 = \sum_{l=1}^{n} \tilde{\boldsymbol{w}}_l \tilde{\boldsymbol{w}}_l^T / n$ and

$$t_{ij} = C\sqrt{\tilde{\theta}_{ij}} \left(d^{-1/2} + \sqrt{n^{-1} \log d} \right) \text{ for all } i \neq j \tag{6}$$

with $\tilde{\theta}_{ij} = n^{-1} \sum_{s=1}^{n} (\tilde{w}_{is}\tilde{w}_{js} - n^{-1} \sum_{t=1}^{n} \tilde{w}_{it}\tilde{w}_{jt})^2$. We denote the eigen-decomposition of $T(\widetilde{\mathbf{\Sigma}}_2)$ by

$$T(\widetilde{\mathbf{\Sigma}}_2) = \sum_{j=1}^{d} \acute{\lambda}_j \acute{\boldsymbol{h}}_j \acute{\boldsymbol{h}}_j^T,$$

where $\acute{\lambda}_j$s are eigenvalues of $T(\widetilde{\mathbf{\Sigma}}_2)$ having $\acute{\lambda}_1 \geq \cdots \geq \acute{\lambda}_d \geq 0$ and $\acute{\boldsymbol{h}}_j$ is a unit eigenvector corresponding to the $\acute{\lambda}_j$. Note that

$$\mathbf{\Sigma}^{-1} = \sum_{j=1}^{m} \lambda_j^{-1} \boldsymbol{h}_j \boldsymbol{h}_j^T + \sum_{j=m+1}^{d} \lambda_j^{-1} \boldsymbol{h}_j \boldsymbol{h}_j^T.$$

Finally, by applying the NR method to the signal part and the noise part, we propose to estimate $\mathbf{\Sigma}^{-1}$ by

$$\widetilde{\mathbf{\Sigma}}^{-1} = \sum_{j=1}^{m} \tilde{\lambda}_j^{-1} \tilde{\boldsymbol{h}}_j \tilde{\boldsymbol{h}}_j^T + \left(\boldsymbol{I}_d - \sum_{j=1}^{m} \tilde{\boldsymbol{h}}_j \tilde{\boldsymbol{h}}_j^T \right) \left(\sum_{j=1}^{d-m} \acute{\lambda}_j^{-1} \acute{\boldsymbol{h}}_j \acute{\boldsymbol{h}}_j^T \right) \left(\boldsymbol{I}_d - \sum_{j=1}^{m} \tilde{\boldsymbol{h}}_j \tilde{\boldsymbol{h}}_j^T \right). \tag{7}$$

Note that the signal and noise parts are orthogonal. We call this new estimation method the "NR-POET". In the next section, we compare the performance of the NR-POET with the POET by several simulations.

5 Simulation studies

In this section, we compare the performance of the NR-POET with the POET and S-POET by using computer simulations. We considered the following covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Gamma}_{d_{(1)},\sigma_{(1)}} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \mathbf{\Gamma}_{d_{(2)},\sigma_{(2)}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \mathbf{\Omega}_{d_{(3)}}(\rho) \end{pmatrix},$$

where $d_{(1)} + d_{(2)} + d_{(3)} = d$, $\mathbf{\Gamma}_{t,\sigma} = \sigma(\boldsymbol{I}_t + \mathbf{1}_t\mathbf{1}_t^T)$ with $\mathbf{1}_t = (1,\ldots,1)^T$ and $\mathbf{\Omega}_t(\rho) = \boldsymbol{B}_t(\rho^{|i-j|^{1/3}})\boldsymbol{B}_t$ with $\boldsymbol{B}_t = \mathrm{diag}[\{0.5 + 1/(t+1)\}^{1/2},\ldots,\{0.5 + t/(t+1)\}^{1/2}]$. Note that $\lambda_{\max}(\mathbf{\Gamma}_{t,\sigma}) = \sigma(t+1)$ and its other eigenvalues are σ. Also, note that $[\lambda_{\max}\{\mathbf{\Omega}_t(\rho)\}]^2/\mathrm{tr}[\{\mathbf{\Omega}_t(\rho)\}^2] = o(1)$ as $t \to \infty$ for $|\rho| < 1$. We set $(d_{(1)}, d_{(2)}) = (\lceil d/3 \rceil, \lceil d/3 \rceil)$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. We set $(\sigma_{(1)}, \sigma_{(2)}) = (1, d^{-1/3})$. Note that $(\lambda_1, \lambda_2) \approx (d/3, d^{2/3}/3)$, so that (C-ii) is satisfied while (C-ii') is not. We set $d = 200(200)1400$, $n = \lceil d^{3/5} \rceil$ and $\rho = 0.3$.

We fixed $C = 5$ in (2) and (6). We considered two cases:

(S-i) $\boldsymbol{x}_i$s are generated from $N_d(\mathbf{0}, \mathbf{\Sigma})$;

(S-ii) $\boldsymbol{x}_i$s are generated from $z_{ri} = (y_{ri} - 2)/2$ $(r = 1,\ldots,d)$ in which y_{ri}s are i.i.d as the chi-squared distribution with 2 degrees of freedom.

For each case, we estimated $\mathbf{\Sigma}^{-1}$ by using the POET in (3), S-POET in (5) and NR-POET in (7). We calculated the average loss and its standard deviation by 2000 (= R, say) times replications for each estimator. Under a fixed scenario, let $\widehat{\mathbf{\Sigma}}_r^{-1}$ be the r-th estimation of $\mathbf{\Sigma}^{-1}$. Let $||\boldsymbol{M}||_\Sigma = ||\mathbf{\Sigma}^{1/2}\boldsymbol{M}\mathbf{\Sigma}^{1/2}||_F/\sqrt{d}$. In the above simulations, $E||\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}||_\Sigma$ was estimated by $R^{-1}\sum_{r=1}^R ||\widehat{\mathbf{\Sigma}}_r^{-1} - \mathbf{\Sigma}^{-1}||_\Sigma$. We denote the standard deviation of $||\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}||_\Sigma$ by $SD||\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}||_\Sigma$. In the above simulations, $SD||\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}||_\Sigma$ was estimated by the standard deviation of $||\widehat{\mathbf{\Sigma}}_r^{-1} - \mathbf{\Sigma}^{-1}||_\Sigma$, $r = 1,\ldots,R$. We displayed the results for (S-i) in Figure 1 and for (S-ii) in Figure 2. We observed that the NR-POET gave better performances than the POET and S-POET both for (S-i) and (S-ii).

Acknowledgements

Research of the second author was partially supported by Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science (JSPS), under Contract Number 18K03409. Research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 15H01678 and 19K22837. This work was supported by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University.

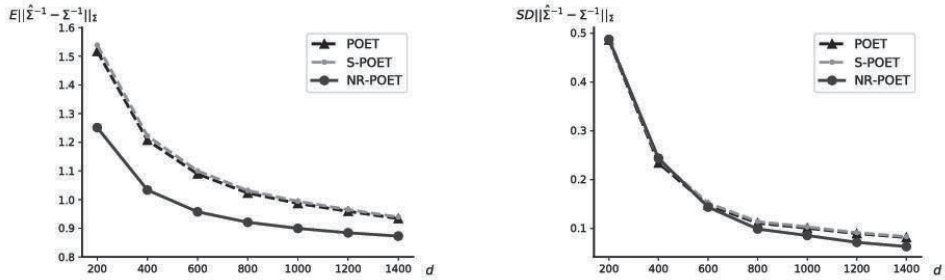

Figure 1: For Case (S-i), the left panel displays the average loss and the right panel displays its standard deviation.

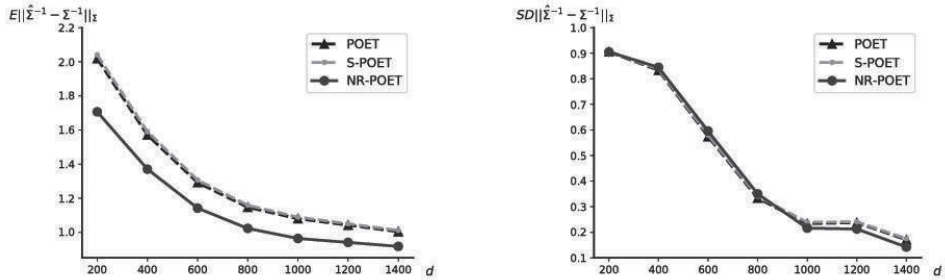

Figure 2: For Case (S-ii), the left panel displays the average loss and the right panel displays its standard deviation.

References

[1] Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, **28**, 43-62.

[2] Aoshima, M. and Yata, K. (2019a). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics*, **71**, 473-503.

[3] Aoshima, M. and Yata, K. (2019b). High-dimensional quadratic classifiers in non-sparse settings. *Methodology and Computing in Applied Probability*, **21**, 663-682.

[4] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, **36**, 2577-2604.

[5] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society, Series B.*, **75**, 603-80.

[6] Jung, S. and Marron, J.S. (2009). PCA consistency in high dimension, low sample size context. *Annals of Statistics*, **37**, 4104-4130.

[7] Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Annals of Statistics*, **45**, 1342-1374.

[8] Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, **105**, 193-215.

[9] Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis*, **122**, 334-354.