Journal of Multivariate Analysis 120 (2013) 102–119

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Asymptotic error bounds for kernel-based Nyström low-rank approximation matrices

Lo-Bin Chang [a,*], Zhidong Bai [b,c], Su-Yun Huang [d], Chii-Ruey Hwang [e]

[a] Department of Applied Mathematics, National Chiao Tung University, Taiwan, ROC
[b] KLASMOE & School of Mathematics and Statistics, Northeast Normal University, China
[c] Department of Statistics and Applied Probability, National University of Singapore, Singapore
[d] Institute of Statistical Science, Academia Sinica, Taiwan, ROC
[e] Institute of Mathematics, Academia Sinica, Taiwan, ROC

HIGHLIGHTS

- Many kernel-based learning algorithms have the computational load.
- The Nyström low-rank approximation is designed for reducing the computation.
- We propose the spectrum decomposition condition with a theoretical justification.
- Asymptotic error bounds on eigenvalues and eigenvectors are derived.
- Numerical experiments are provided for covariance kernel and Wishart matrix.

ARTICLE INFO

Article history:
Received 3 April 2012
Available online 28 May 2013

AMS subject classifications:
60F99
62H30
68T10
68W25

Keywords:
Nyström approximation
Kernel Gram matrix
Spectrum decomposition
Asymptotic error bound
Wishart random matrix

ABSTRACT

Many kernel-based learning algorithms have the computational load scaled with the sample size n due to the column size of a full kernel Gram matrix $\boldsymbol{K}$. This article considers the Nyström low-rank approximation. It uses a reduced kernel $\widehat{\boldsymbol{K}}$, which is $n \times m$, consisting of m columns (say columns $i_1, i_2, \cdots, i_m$) randomly drawn from $\boldsymbol{K}$. This approximation takes the form $\boldsymbol{K} \approx \widehat{\boldsymbol{K}}\boldsymbol{U}^{-1}\widehat{\boldsymbol{K}}^T$, where $\boldsymbol{U}$ is the reduced $m \times m$ matrix formed by rows $i_1, i_2, \cdots, i_m$ of $\widehat{\boldsymbol{K}}$. Often m is much smaller than the sample size n resulting in a thin rectangular reduced kernel, and it leads to learning algorithms scaled with the column size m. The quality of matrix approximations can be assessed by the closeness of their eigenvalues and eigenvectors. In this article, asymptotic error bounds on eigenvalues and eigenvectors are derived for the Nyström low-rank approximation matrix.

1. Introduction

Due to the fast advancement of information technology, kernel-based learning algorithms have become popular nowadays and they play an important role in machine learning with ample applications in statistics, biostatistics, medical science, image analysis, pattern recognition, engineering, etc. (See, e.g., [4,5,20,1,12].) Kernel functions are flexible building blocks for modeling complex and nonlinear data structures. The value of a kernel function $K(x, y)$ represents a dot product in a kernel-induced Hilbert space, often high-dimensional or even infinite dimensional, and can be interpreted as the similarity measure between the two points, x and y.

* Correspondence to: Department of Applied Mathematics, National Chiao Tung University, Hsin-Chu 30010, Taiwan, ROC.
E-mail addresses: r91221024@yahoo.com.tw, lobin_chang@math.nctu.edu.tw (L.-B. Chang).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.05.006

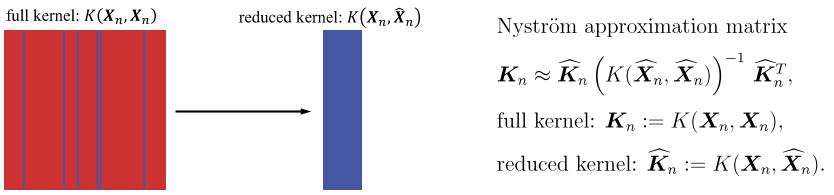

Fig. 1. Reduced kernel and Nyström approximation.

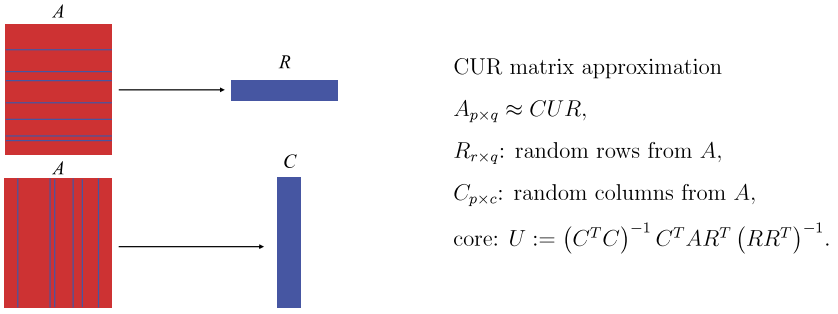

Fig. 2. CUR decomposition.

Many kernel-based learning algorithms have computational load scaled with the sample size n of data collection $\{X_1, \ldots, X_n\}$. This article considers the Nyström low-rank approximation to kernel-based Gram matrix and provides it with a theoretical justification. Asymptotic error bounds on eigenvalues and eigenvectors are derived for the Nyström low-rank approximation matrix. The lack of intuition for the approximation of eigenvectors creates a great surprise on numerical results of asymptotic error. However, to our best knowledge, there is no other existing article mathematically exploring the convergence or error bound of eigenvectors for Nyström approximation matrix. The Nyström method is an easy and yet efficient approach for low-rank approximation, which dramatically cuts down the computational load and memory usage. See, for instance, Lee and Mangasarian [16], Williams and Seeger [23], Drineas and Mahoney [9], Lee and Huang [15] for studies of Nyström low-rank approximation for kernel matrices.

The underlying kernel function $K(x, y)$ in this article is assumed continuous, symmetric, nonnegative definite, and defined on $\mathcal{X} \times \mathcal{X}$. Let X be a random variable having continuous distribution F on $\mathcal{X} \subset \Re^p$. Let $\boldsymbol{X}_n$ be the data matrix (random) consisting of i.i.d. copies of X, i.e., $\boldsymbol{X}_n := (X_1, \ldots, X_n)^T$, which is of size $n \times p$; and let $\boldsymbol{K}_n := K(\boldsymbol{X}_n, \boldsymbol{X}_n) = [K(X_i, X_j)]_{i,j=1}^n$ be the full kernel matrix. The key idea of Nyström approximation is to employ a reduced kernel. It randomly selects a portion of data set to generate a thin rectangular kernel matrix, called reduced kernel and denoted by $\widehat{\boldsymbol{K}}_n := K(\boldsymbol{X}_n, \widehat{\boldsymbol{X}}_n)$, where $\widehat{\boldsymbol{X}}_n$ is a data subset matrix formed by a subset of $\{X_1, \ldots, X_n\}$. Then, it uses this much smaller rectangular kernel matrix to replace or to generate an approximation to the full kernel matrix. See Fig. 1 for graphical illustration.

The technique of using a reduced kernel matrix has been successfully applied to other kernel-based learning algorithms, such as least squares support vector machine [21,22], proximal support vector machine [11], Lagrangian support vector machine [18], active set support vector regression [19], smooth ϵ-support vector regression [14], kernel sliced-inverse regression [24] and robust kernel PCA [13], among others.

The random subsample $\{K(\cdot, X_{i_k})\}_{k=1}^m$ is used as a basis subset to replace the full-sample basis set $\{K(\cdot, X_i)\}_{i=1}^n$. In a training phase of a kernel algorithm, the thin reduced kernel matrix $\boldsymbol{K}_n = K(\boldsymbol{X}_n, \widehat{\boldsymbol{X}}_n)$ is used as data inputs, where $\widehat{\boldsymbol{X}}_n$ consists of $\{X_{i_k}\}_{k=1}^m$. Notice that the number of observations (the column size of $\boldsymbol{K}_n$) is not reduced, it is the number of basis functions (the row size of $\boldsymbol{K}_n$) that has been cut down. This uniform random subset for kernel basis selection has a link to the popular uniform design, which is a space filling design. Space filling designs are known to be robust against the worst possible scenario [10]. Of course, there is always a random luck issue in every random sampling scheme. To improve the quality of the random subsample used as partial kernel basis, a stratified random subset is suggested. For classification problem, the random sampling has to be stratified over classes. For regression problem, the random sampling has to be stratified over the regression responses. Furthermore, the low-rank approximation matrix actually adopts a model with less model complexity, thus a larger penalty is suggested to enforce better data fidelity. See Lee and Huang [15] for more detailed discussion and suggestions for practical implementation.

The idea of using a random subset can also be found in a series of works of CUR matrix decompositions ([6–8,17], and references therein). See Fig. 2 for illustration of a CUR decomposition.

In addition to being continuous, symmetric and nonnegative definite, the kernel function K is assumed square-integrable

$$\int_{\mathcal{X}\times\mathcal{X}} K^2(x, y)dF(x)dF(y) = c < \infty, \tag{1}$$

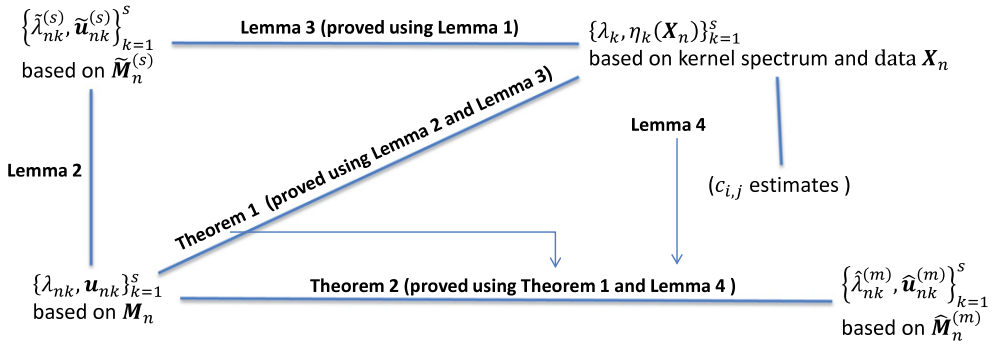

Fig. 3. Connection diagram. The correspondences among variables, lemmas and theorems can be viewed in this diagram. The major goal is to connect $\{\lambda_{nk}, \boldsymbol{u}_{nk}\}$ (with full kernel matrix) to $\{\widehat{\lambda}_{nk}^{(m)}, \widehat{\boldsymbol{u}}_{nk}^{(m)}\}$ (with reduced-rank approximation matrix).

has the following spectrum decomposition

$$K(x, y) = \sum_{k=1}^{\infty} \lambda_k \eta_k(x) \eta_k(y), \quad \text{where } \int_{\mathcal{X}} \eta_k(x) \eta_j(x) dF(x) = \delta_{kj}, \tag{2}$$

and is of trace type, i.e.,

$$\sum_{k=1}^{\infty} \lambda_k < \infty. \tag{3}$$

For simplicity, we assume eigenvalues of K are strictly positive, distinct, and arranged in descending order

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_k > \cdots > 0. \tag{4}$$

(Notice that the method is designed to work for symmetric nonnegative definite kernels, not for others (see the second example in Section 4). The number of positive eigenvalues of the kernel needs to be greater than n for $n \times n$ data kernel matrices defined below. If the eigenvalues are not distinct, the corresponding eigenspaces of non-distinct eigenvalues are of more than one dimension. Thus, the convergence result for eigenvalues still holds, but the convergence for eigenvectors need to be modified in terms of eigenspaces.) Consider the $n \times n$ kernel data matrix (scaled by n^{-1})

$$\boldsymbol{M}_n := n^{-1} K(\boldsymbol{X}_n, \boldsymbol{X}_n) = n^{-1} \left[K(X_i, X_j)\right]_{i,j=1}^{n} = n^{-1} \boldsymbol{K}_n \tag{5}$$

with eigenvalues $\lambda_{n1} \geq \lambda_{n2} \geq \lambda_{n3} \geq \cdots \geq \lambda_{nn} \geq 0$ and corresponding unit eigenvectors $\boldsymbol{u}_{nk}, k = 1, 2, \ldots, n$. Matrices $\boldsymbol{K}_n$ and $\boldsymbol{M}_n$ are both called a full kernel matrix. The eigenvalue decomposition problem for a full kernel matrix $\boldsymbol{M}_n$ is computationally costly. An alternative is to resort to a reduced kernel by random subset. Since data are i.i.d. copies of X, without loss of generality, we may assume that the random subset, denoted by $\widehat{\boldsymbol{X}}_n^{(m)}$, is formed by $\{X_1, \ldots, X_m\}$ for some $m < n$. Consider the partition of $\boldsymbol{M}_n$ as

$$\boldsymbol{M}_n = \begin{bmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{bmatrix} \begin{matrix} m \\ n-m \end{matrix} \quad . \tag{6}$$
$$\begin{matrix} m & n-m \end{matrix}$$

The following rank-m approximation matrix is called Nyström approximation to $\boldsymbol{M}_n$

$$\widehat{\boldsymbol{M}}_n^{(m)} := \begin{bmatrix} \boldsymbol{M}_{11} \\ \boldsymbol{M}_{21} \end{bmatrix} \boldsymbol{M}_{11}^{-1} \begin{bmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{21} \boldsymbol{M}_{11}^{-1} \boldsymbol{M}_{12} \end{bmatrix}. \tag{7}$$

This approximation is based on a reduced kernel $\frac{1}{n} K(\boldsymbol{X}_n, \widehat{\boldsymbol{X}}_n^{(m)}) = \left[\boldsymbol{M}_{11}^T, \boldsymbol{M}_{21}^T\right]^T$, which is of size $n \times m$. Denote the kth eigenvalue and its associated eigenvector of $\widehat{\boldsymbol{M}}_n^{(m)}$ by $\widehat{\lambda}_{nk}^{(m)}$ and $\widehat{\boldsymbol{u}}_{nk}^{(m)}$. The aims of this article are

(i) to study the asymptotic orders of magnitude for the full kernel $\{\lambda_{nk}, \boldsymbol{u}_{nk}\}$ and the reduced kernel $\{\widehat{\lambda}_{nk}^{(m)}, \widehat{\boldsymbol{u}}_{nk}^{(m)}\}$, as compared to the ideal ones $\left\{\lambda_k, \frac{1}{\sqrt{n}} \eta_k(\boldsymbol{X}_n)\right\}$, where $\eta_k(\boldsymbol{X}_n)$ is an n-vector given by $\eta_k(\boldsymbol{X}_n) := \left(\eta_k(X_1), \ldots, \eta_k(X_n)\right)^T$; and
(ii) to find the asymptotic bounds for $\boldsymbol{M}_n - \widehat{\boldsymbol{M}}_n^{(m)}$ in terms of their eigenvalues and eigenvectors, where $\boldsymbol{M}_n - \widehat{\boldsymbol{M}}_n^{(m)}$ is the difference between the full kernel matrix and the reduced-rank approximation matrix.

Results established in this article are summarized in Fig. 3. The $\widetilde{\lambda}_{nk}^{(s)}$, $\widetilde{\boldsymbol{u}}_{nk}^{(s)}$ and $\widetilde{\boldsymbol{M}}_n^{(s)}$ in the diagram are intermediate variables defined later in the paragraph of Eq. (17) to prove our main results. Notice that, two opposite directions can be chosen for an

eigenvector. For convenience, the directions of the eigenvectors, $\boldsymbol{u}_{nk}$, $\widehat{\boldsymbol{u}}_{nk}^{(m)}$, $\widetilde{\boldsymbol{u}}_{nk}^{(s)}$ are chosen toward the direction of $\eta_k(\boldsymbol{X}_n)$ for each k. The rest of the article is organized as follows. Main results of eigenvalues and eigenvectors error bounds for Nyström approximation matrix are given in Section 2. Technical lemmas and proofs are placed in Section 3. Two numerical examples are displayed in Section 4. A list of notation usage is appended at the end of the article.

2. Main results

General assumptions. Throughout the rest of this article, we assume that X is a random variable with continuous distribution F on $\mathcal{X} \subset \Re^p$, and that the kernel function K is continuous, symmetric, nonnegative definite and satisfies conditions (1)–(4). Further assume that $EK^{2+\tau}(X, X) < \infty$ for some $\tau > 0$. (For bounded kernel, e.g., Gaussian kernel, τ can be any positive number.)

Theorem 1. *Let $s = n^\alpha$ with $0 < \alpha < \frac{\tau}{2(4+2\tau)}$. Then, for any α, ν with $\alpha < \nu < \frac{\tau}{4+2\tau} - \alpha$, we have*

$$\sum_{k=1}^{s} (\lambda_k - \lambda_{nk})^2 = O_p(\epsilon_{n,s}) + O_p\left(sn^{-\nu}\right), \tag{8}$$

where

$$\epsilon_{n,s} = \frac{1}{n}\left(\sum_{k=s+1}^{\infty} \lambda_k\right)^2 + \sum_{k=s+1}^{\infty} \lambda_k^2.$$

Furthermore, for any fixed k, we have

$$\lambda_{nk} = \lambda_k + O_p\left(\sqrt{\epsilon_{n,s}}\right) + O_p\left(\sqrt{sn^{-\nu}}\right), \tag{9}$$

$$\boldsymbol{u}_{nk} = \frac{1}{\sqrt{n}}\eta_k(\boldsymbol{X}_n) + O_p^{(L_2)}\left(\sqrt{\epsilon_{n,s}}\right) + O_p^{(L_2)}\left(\sqrt{sn^{-2\nu}}\right). \tag{10}$$

Remark 1. For a bounded kernel K, we can set τ to be any fixed but arbitrarily large number. Since $0 < \nu - \alpha < \frac{\tau}{4+2\tau} - 2\alpha$, when $\tau \to \infty$, we can take $\nu - \alpha$ to be very close to $\frac{1}{2} - 2\alpha$ $\left(\text{say } \nu - \alpha = \frac{1}{2} - 2\alpha - \delta \text{ for small positive } \delta\right)$. Thus, $\sqrt{sn^{-\nu}} = n^{-\frac{1}{2}(\nu-\alpha)} = n^{-\frac{1}{4}+\alpha+\frac{1}{2}\delta}$, and $\sqrt{sn^{-2\nu}} = n^{-\nu+\frac{\alpha}{2}} = n^{-\frac{1}{2}+\frac{3\alpha}{2}+\delta}$. Assume that λ_k has a polynomial decay, i.e., $\lambda_k = O(k^{-\beta})$ for some $\beta > 1$. Then,

$$\epsilon_{n,s} = \frac{1}{n}\left(\sum_{k=s+1}^{\infty} \lambda_k\right)^2 + \sum_{k=s+1}^{\infty} \lambda_k^2 = O\left(s^{-2\beta+1}\right) = O\left(n^{-\alpha(2\beta-1)}\right).$$

To balance the order of $\sqrt{\epsilon_{n,s}}$ and $\sqrt{sn^{-\nu}}$, we take $\alpha = \frac{1-2\delta}{4\beta+2}$. Then,

$$\sqrt{\epsilon_{n,s}} = O\left(\sqrt{sn^{-\nu}}\right) = O\left(n^{-\frac{1}{4}+\frac{1-2\delta}{4\beta+2}+\frac{\delta}{2}}\right)$$

which is very close to $O\left(n^{-\frac{1}{4}+\frac{1}{4\beta+2}}\right)$ when δ is close to 0. If the underlying kernel K has a faster eigenvalue decay (larger β or an exponential decay), then it leads to faster convergence rates for λ_{nk} and $\boldsymbol{u}_{nk}$.

Theorem 2. *Let $s_m = m^\alpha$ with $0 < \alpha < \frac{\tau}{2(4+2\tau)}$, and let*

$$\epsilon_{m,s_m} = \frac{1}{m}\left(\sum_{k=s_m+1}^{\infty} \lambda_k\right)^2 + \sum_{k=s_m+1}^{\infty} \lambda_k^2$$

$$\tilde{\epsilon}_{m,n} = \left(\frac{n-m}{n}\right)^2 \left[\left(\sum_{r=1}^{l_m} \frac{1}{\lambda_r}\right)^2 \left(\epsilon_{m,s_m} + s_m m^{-\nu}\right) + \sum_{k=l_m+1}^{\infty} \lambda_k\right] + \frac{n-m}{n^2},$$

where $\{l_m\}_{m=1}^{\infty}$ is an integer sequence satisfying

$$\lim_{m\to\infty} \left[\left(\sum_{r=1}^{l_m} \frac{1}{\lambda_r}\right)^2 \left(\epsilon_{m,s_m} + s_m m^{-\nu}\right) + \sum_{k=l_m+1}^{\infty} \lambda_k\right] = 0. \tag{11}$$

Then, for any α, ν *with* $\alpha < \nu < \frac{\tau}{4+2\tau} - \alpha$, *we have*

$$\sum_{k=1}^{n} \left(\lambda_{nk} - \widehat{\lambda}_{nk}^{(m)}\right)^2 = O_p(\tilde{\epsilon}_{m,n}). \tag{12}$$

Furthermore, for any fixed k, we have

$$\widehat{\lambda}_{nk}^{(m)} = \lambda_{nk} + O_p\left(\sqrt{\tilde{\epsilon}_{m,n}}\right), \tag{13}$$

$$\widehat{\boldsymbol{u}}_{nk}^{(m)} = \boldsymbol{u}_{nk} + O_p^{(L_2)}\left(\sqrt{\tilde{\epsilon}_{m,n}}\right). \tag{14}$$

Remark 2. The $\frac{1}{\lambda_r}$'s in Eq. (11) come from $\boldsymbol{M}_{11}^{-1}$ (in Eq. (7) for the Nyström approximation). They amplify the error term, $\epsilon_{m,s_m} + s_m m^{-\nu}$ in our estimation approach, so smaller l_m makes better convergence for the first term of Eq. (11). However, as m goes to infinity, making the second term $\sum_{k=l_m+1}^{\infty} \lambda_k$ convergent faster to zero requires that l_m tends to infinity with faster rate. Thus, to obtain the optimal convergence rate for $\tilde{\epsilon}_{m,n}$, we choose l_m such that

$$\left(\sum_{r=1}^{l_m} \frac{1}{\lambda_r}\right)^2 \left(\epsilon_{m,s_m} + s_m m^{-\nu}\right) = O\left(\sum_{k=l_m+1}^{\infty} \lambda_k\right)$$

(see the example of next remark).

Remark 3. Assume that λ_k has a polynomial decay, i.e., $\lambda_k = O(k^{-\beta})$ for some $\beta > 1$. Then, $\tilde{\epsilon}_{m,n} = O\big(l_m^{2\beta+2} \times (s_m^{-2\beta+1} + s_m m^{-\nu}) + l_m^{-\beta+1}\big)$. Similar to Remark 1, when τ can be arbitrarily large, by taking $\nu - \alpha = \frac{1}{2} - 2\alpha - \delta$ and $\alpha = \frac{1-2\delta}{4\beta+2}$ for some small positive δ, we can get $s_m m^{-\nu} = O\left(s_m^{-2\beta+1}\right)$. Therefore, $\tilde{\epsilon}_{m,n} = O\left(l_m^{2\beta+2} s_m^{-2\beta+1} + l_m^{-\beta+1}\right)$. Now let l_m be the largest integer less than or equal to s_m^b. Again to balance $l_m^{2\beta+2} s_m^{-2\beta+1}$ and $l_m^{-\beta+1}$, we take $b = \frac{2\beta-1}{3\beta+1}$. Thus, we get

$$\tilde{\epsilon}_{m,n} = O\left(m^{-\frac{1}{2}\frac{(\beta-1)(2\beta-1)}{(3\beta+1)(2\beta+1)}(1-2\delta)}\right).$$

When $\beta \to \infty$, the convergence rate can be as close to the rate of $O\left(m^{-\frac{1}{6}}\right)$ as possible by taking δ close to 0.

Remark 4. In practice, the size of the subsample m is chosen to be much smaller than the sample size n, but according to the rate of $\tilde{\epsilon}_{m,n}$, even for the fastest rate of $O\left(m^{-\frac{1}{6}}\right)$ discussed in Remark 3, m has to be chosen big enough to make the approximation error small. From the numerical experiments in Section 4, the convergence rate, in fact, can be faster than the upper bound rate $O(\tilde{\epsilon}_{m,n})$. Thus, this upper bound is not tight. Nevertheless, our results in Theorem 2 are enough to show that the errors converge to zero as both n, m tend to infinity. We will further discuss the issue below in technical lemmas. Also to our best knowledge, there is no other existing article mathematically exploring the convergence or error bound of *eigenvectors* for Nyström approximation.

3. Technical lemmas

Before moving into the mathematical details in this section, let us overview the connection between the following four lemmas and two theorems. As in Fig. 3, Theorem 1 will be proved using Lemma 2 and Lemma 3, and Lemma 1 will be used to prove Lemma 3. To prove Theorem 2 we will need Lemma 4 together with Theorem 1.

All the technical lemmas are derived under the *General assumptions* stated in Section 2. Define

$$\boldsymbol{V}_n^{(s)} := \frac{1}{\sqrt{n}}\big(\eta_1(\boldsymbol{X_n}), \eta_2(\boldsymbol{X_n}), \ldots, \eta_s(\boldsymbol{X_n})\big), \tag{15}$$

which is an $n \times s$ matrix, and define $\boldsymbol{A}_n^{(s)} := \boldsymbol{V}_n^{(s)T} \boldsymbol{V}_n^{(s)} := \left[a_{ij}\right]_{i,j=1}^{s}$. We have the following lemma, which basically says that these n-vectors $\{n^{-1/2}\eta_k(\boldsymbol{X}_n)\}_{k=1}^{s}$ are nearly orthonormal, when n is large.

Lemma 1. *Let* $s = n^{\alpha}$ *with* $0 < \alpha < \frac{\tau}{4+2\tau}$. *Then, for any* $0 < \nu < \frac{\tau}{4+2\tau} - \alpha$, *we have*

$$\boldsymbol{A}_n^{(s)} = \boldsymbol{I}_s + O_p^{(\infty)}\left(n^{-\nu}\right),$$

where the order $O_p^{(\infty)}(\cdot)$ is in the sense of being in probability and under the entrywise L_∞ matrix norm, i.e., in uniform sense. Moreover, if $s = n^\alpha$ with $0 < \alpha < \frac{\tau}{2(4+2\tau)}$, then for any $\alpha < \nu < \frac{\tau}{4+2\tau} - \alpha$, we have

$$\left\{\boldsymbol{A}_n^{(s)}\right\}^{-1/2} = \boldsymbol{I}_s + O_p^{(\infty)}\left(sn^{-\nu}\right) = \boldsymbol{I}_s + O_p^{(\infty)}\left(n^{-(\nu-\alpha)}\right).$$

(Note that $\boldsymbol{A}_n^{(s)}$ has faster rate than $\{\boldsymbol{A}_n^{(s)}\}^{-1/2}$ in convergence to $\boldsymbol{I}_s$.)

Proof. The (k, l)th entry in $\boldsymbol{A}_n^{(s)}$ is given by $a_{kl} = \frac{1}{n}\eta_k(\boldsymbol{X}_n)^T\eta_l(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^n \eta_k(X_i)\eta_l(X_i)$. For a fixed $\varsigma > 0$, consider the truncation

$$Y_{ki} = \begin{cases} \eta_k(X_i), & \text{if } |\eta_k(X_i)| < \varsigma n^{1/(2+\tau)}, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\widehat{a}_{kl} = \frac{1}{n}\sum_{i=1}^n Y_{ki}Y_{li}$ and $Z_{kl}(X_i) = |Y_{ki}Y_{li} - E(Y_{ki}Y_{li})|$. Note that, for $j \geq 2$, we have

$$E\{Z_{kl}^j(X)\} \leq \mathrm{var}\{Z_{kl}(X)\}\left(\varsigma^2 n^{2/(2+\tau)}\right)^{j-2} = O\left((\varsigma^2 n^{2/(2+\tau)})^{j-2}\right), \tag{16}$$

and for $j < 2$, we have $E\{Z_{kl}^j(X)\} = O(1)$. Since $E|\eta_k(X)|^{4+2\tau} < \infty$, we have for any fixed $\varsigma > 0$,

$$\begin{aligned} P\left\{\eta_k(X_i) \neq Y_{ki},\ \forall k, i \leq n\right\} &\leq n^2 P\left\{|\eta_k(X)| \geq \varsigma n^{1/(2+\tau)}\right\} \\ &\leq \varsigma^{-(4+2\tau)} E\left\{|\eta_k(X)|^{4+2\tau} I\left(|\eta_k(X)| \geq \varsigma n^{1/(2+\tau)}\right)\right\} \to 0, \quad \text{as } n \to \infty. \end{aligned}$$

Thus, $n^2 P\{\eta_k(X_i) \neq Y_{ki}\} \to 0$ for any fixed $k, i \leq n$.

Let δ_{kl} be the Kronecker delta. Assume that s is even (otherwise replace s by $s-1$). We have, for $\varsigma, \varepsilon > 0$ and small enough,

$$\begin{aligned} P\left\{\max_{k,l\leq s}|a_{kl} - \delta_{kl}| > \varepsilon\right\} &\leq P\left\{\max_{k,l\leq s}|\widehat{a}_{kl} - \delta_{kl}| > \varepsilon\right\} + P\left\{\eta_k(X_i) \neq Y_{ki}, k, i \leq n\right\} \\ &\leq \sum_{k,l=1}^s P\left\{\frac{1}{n}\left|\sum_{i=1}^n (Y_{ki}Y_{li} - \delta_{kl})\right| > \varepsilon\right\} + o(1) \\ &\leq \sum_{k,l=1}^s P\left\{\frac{1}{n}\left|\sum_{i=1}^n (Y_{ki}Y_{li} - E(Y_{ki}Y_{li}))\right| \geq \frac{\varepsilon}{2}\right\} + P\left\{\frac{1}{n}\left|\sum_{i=1}^n (E(Y_{ki}Y_{li}) - \delta_{kl})\right| > \frac{\varepsilon}{2}\right\} + o(1) \\ &\leq \sum_{k,l=1}^s \left(\frac{n\varepsilon}{2}\right)^{-s} E\left|\sum_{i=1}^n \left(Y_{ki}Y_{li} - E(Y_{ki}Y_{li})\right)\right|^s + P\left\{|E(Y_{k1}Y_{l1}) - \delta_{kl}| > \frac{\varepsilon}{2}\right\} + o(1), \end{aligned}$$

Now, when n is large enough, $|E(Y_{k1}Y_{l1}) - \delta_{kl}| < \frac{\varepsilon}{2}$. Therefore, the second term above is zero; and since s is even, we have

$$\begin{aligned} P\left\{\max_{k,l\leq s}|a_{kl} - \delta_{kl}| > \varepsilon\right\} &\leq \left(\frac{n\varepsilon}{2}\right)^{-s} \sum_{k,l=1}^s E\left(\sum_{i=1}^n \left(Y_{ki}Y_{li} - E(Y_{ki}Y_{li})\right)\right)^s + 0 + o(1) \\ &= \left(\frac{n\varepsilon}{2}\right)^{-s} \sum_{k,l=1}^s \sum_{1\leq r\leq s} \sum_{1\leq j_1<\cdots<j_r\leq n} \sum_{\substack{i_1+\cdots+i_r=s \\ i_1,\ldots,i_r\geq 1}} \frac{s!\prod_{t=1}^r E\left(Y_{ki}Y_{li} - E(Y_{ki}Y_{li})\right)^{i_t}}{i_1!\cdots i_r!} + o(1). \end{aligned}$$

Since $E\left(Y_{ki}Y_{li} - E(Y_{ki}Y_{li})\right) = 0$, we can only consider the indices with $i_t \geq 2$ for $i = 1, 2, \ldots, r$. Thus the index r is now considered to be less than $s/2$. Using the fact $E\left(Y_{ki}Y_{li} - E(Y_{ki}Y_{li})\right)^{i_t} \leq E\{Z_{kl}^{i_t}(X_i)\}$, for $i_t \geq 2$ and Eq. (16), we obtain

$$\begin{aligned} P\left\{\max_{k,l\leq s}|a_{kl} - \delta_{kl}| > \varepsilon\right\} &\leq \left(\frac{n\varepsilon}{2}\right)^{-s} \sum_{k,l=1}^s \sum_{1\leq r\leq s/2} \sum_{1\leq j_1<\cdots<j_r\leq n} \sum_{\substack{i_1+\cdots+i_r=s \\ i_1,\ldots,i_r\geq 2}} \frac{s!E\{Z_{kl}^{i_1}(X_{j_1})\}\cdots E\{Z_{kl}^{i_r}(X_{j_r})\}}{i_1!\cdots i_r!} + o(1), \\ &\leq \sum_{k,l=1}^s \left(\frac{n\varepsilon}{2}\right)^{-s} \sum_{1\leq r\leq s/2} \frac{n^r}{r!}\cdot r^s(\varsigma^2 n^{2/(2+\tau)})^{s-2r} + o(1) \\ &\leq \sum_{k,l=1}^s \sum_{1\leq r\leq s/2} \left(\frac{2\varsigma^2}{\varepsilon}\right)^s n^{-(s-r)\tau/(2+\tau)} \frac{1}{\varsigma^{4r}}\cdot\frac{r^s}{r!} + o(1) \\ &\leq c_{\varsigma,s}\,\varepsilon^{-s} n^{2-s\tau/(4+2\tau)} + o(1), \end{aligned}$$

where

$$\begin{aligned} c_{\varsigma,s} &= \sum_{1\le r\le s/2} \frac{(2r)^s\,\varsigma^{2s-4r}}{r!} \le \sum_{1\le r\le s/2} \frac{(2r)^s\,\varsigma^{2s-4r}}{e\sqrt{2\pi r}e^{-r}r^r} \le \sum_{1\le r\le s/2} \frac{2^s\varsigma^{2s}}{e\sqrt{2\pi}} s^s (e/\varsigma^4)^r \\ &\le \kappa a^s s^s, \quad \text{for some constants } \kappa \text{ and } a \\ &= \kappa a^s n^{\alpha s}. \end{aligned}$$

Therefore, by choosing $\varepsilon = bn^{-\nu}$, we have

$$P\left\{\max_{k,l\le s}|a_{kl}-\delta_{kl}| > \varepsilon\right\} \le K(a/b)^s n^{-s(\tau/(4+2\tau)-\alpha-\nu)+2} + o(1) \to 0,$$

as $n\to\infty$, since $s = n^\alpha$ and $\tau/(4+2\tau)-\alpha-\nu > 0$. This gives us the first equation. For the second equation, we will use the equivalent property of $\|\boldsymbol{A}\|_2 = \sup\{\|\boldsymbol{A}\boldsymbol{u}\|_2 : \|\boldsymbol{u}\|_2 = 1\}$ and $\|\boldsymbol{A}\|_\infty = \max_{i,j}|a_{ij}|$, i.e.,

$$\|\boldsymbol{A}\|_\infty \le \|\boldsymbol{A}\|_2 \le s\|\boldsymbol{A}\|_\infty, \quad \text{where } \boldsymbol{A} \text{ is an } s\times s \text{ matrix.}$$

Let $\boldsymbol{A}_n^{(s)} = \boldsymbol{P}^T \mathrm{diag}(\theta_1,\theta_2,\ldots,\theta_s)\boldsymbol{P}$, where $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}_s$. Then,

$$\|\boldsymbol{A}_n^{(s)} - \boldsymbol{I}_s\|_2 = \|\mathrm{diag}(\theta_1,\ldots,\theta_s) - \boldsymbol{I}_s\|_2 = \max_i |\theta_i - 1| \le s\|\boldsymbol{A}_n^{(s)} - \boldsymbol{I}_s\|_\infty,$$

and then

$$\begin{aligned} \left\|\left(\boldsymbol{A}_n^{(s)}\right)^{1/2} - \boldsymbol{I}_s\right\|_\infty &\le \left\|\left(\boldsymbol{A}_n^{(s)}\right)^{1/2} - \boldsymbol{I}_s\right\|_2 = \max_i \left|\sqrt{\theta_i} - 1\right| \\ &\le \max_i |\theta_i - 1| \le s\|\boldsymbol{A}_n^{(s)} - \boldsymbol{I}_s\|_\infty. \end{aligned}$$

Therefore, $\left(\boldsymbol{A}_n^{(s)}\right)^{-1/2} = \boldsymbol{I}_s + O_p^{(\infty)}\left(sn^{-\nu}\right)$. □

For an integer $s \le n$, let $\widetilde{K}^{(s)}(x,y)$ be the truncated kernel at s components and let $\widetilde{\boldsymbol{M}}_n^{(s)}$ be its corresponding kernel data matrix scaled by n, i.e.,

$$\widetilde{K}^{(s)}(x,y) := \sum_{k=1}^{s} \lambda_k \eta_k(x)\eta_k(y), \tag{17}$$

$$\widetilde{\boldsymbol{M}}_n^{(s)} := \frac{1}{n}\left[\widetilde{K}^{(s)}(X_i,X_j)\right]_{i,j=1}^n =: \frac{1}{n}\widetilde{\boldsymbol{K}}_n^{(s)}. \tag{18}$$

Denote the leading s eigenvalues of $\widetilde{\boldsymbol{M}}_n^{(s)}$ by $\widetilde{\lambda}_{n1}^{(s)} \ge \widetilde{\lambda}_{n2}^{(s)} \ge \cdots \ge \widetilde{\lambda}_{ns}^{(s)} \ge 0$ and corresponding unit eigenvectors $\widetilde{\boldsymbol{u}}_{nk}^{(s)}$, $k = 1,2,\ldots,s$. Note that $\widetilde{K}^{(s)}(x,y)$ is the best rank-s approximation, in the sense of minimal L_2-norm, to the underlying kernel function $K(x,y)$.

Lemma 2. *For any $s \le n$, we have*

$$\sum_{k=1}^{n}\left(\lambda_{nk} - \widetilde{\lambda}_{nk}^{(s)}\right)^2 = O_p(\epsilon_{n,s}), \tag{19}$$

where $\widetilde{\lambda}_{nk}^{(s)} = 0$ for $k > s$, and $\epsilon_{n,s} = \frac{1}{n}\left(\sum_{k=s+1}^{\infty}\lambda_k\right)^2 + \sum_{k=s+1}^{\infty}\lambda_k^2$. Furthermore, for any fixed k and $s = n^\alpha$ with $0 < \alpha < \frac{\tau}{2(4+2\tau)}$, we have

$$\widetilde{\lambda}_{nk}^{(s)} = \lambda_{nk} + O_p\left(\sqrt{\epsilon_{n,s}}\right), \tag{20}$$

$$\widetilde{\boldsymbol{u}}_{nk}^{(s)} = \boldsymbol{u}_{nk} + O_p^{L_2}\left(\sqrt{\epsilon_{n,s}}\right) + O_p^{L_2}\left(\sqrt{sn^{-\nu}}\right). \tag{21}$$

Proof. Note that the condition $EK^2(X,X) < \infty$ implies that $E\eta_k^4(X) < \infty$. By Lemma 2.3 of Bai [3], we have

$$\begin{aligned} \sum_{k=1}^{n}\left(\lambda_{nk} - \widetilde{\lambda}_{nk}^{(s)}\right)^2 &\le \mathrm{tr}\left(\boldsymbol{M}_n - \widetilde{\boldsymbol{M}}_n^{(s)}\right)^2 \\ &= \frac{1}{n^2}\sum_{k,k'=s+1}^{\infty}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_k\lambda_{k'}\eta_k(X_i)\eta_{k'}(X_i)\eta_k(X_j)\eta_{k'}(X_j). \end{aligned} \tag{22}$$

Let $\eta_k(X_i)\eta_{k'}(X_i)\eta_k(X_j)\eta_{k'}(X_j) = D$. Then, by simple calculation, we get

$$E(D) = \begin{cases} E\eta_k^4(X_i), & \text{if } k = k',\ i = j, \\ E\eta_k^2(X_i)\eta_k^2(X_j) = 1, & \text{if } k = k',\ i \neq j, \\ E\eta_k^2(X_i)\eta_{k'}^2(X_i) \leq \sqrt{E\eta_k^4(X_i)}\sqrt{E\eta_{k'}^4(X_i)}, & \text{if } k \neq k',\ i = j, \\ E\eta_k(X_i)\eta_{k'}(X_i)E\eta_k(X_j)\eta_{k'}(X_j) = 0, & \text{if } k \neq k',\ i \neq j. \end{cases} \tag{23}$$

Therefore, the expectation of the R.H.S. of (22) can be bounded by

$$\frac{1}{n}\sum_{k=s+1}^{\infty}\lambda_k^2 E\eta_k^4(X) + \frac{n-1}{n}\sum_{k=s+1}^{\infty}\lambda_k^2 + \frac{1}{n}\sum_{\substack{k,k'=s+1 \\ k\neq k'}}^{\infty}\lambda_k\lambda_{k'}\sqrt{E\eta_k^4(X)}\sqrt{E\eta_{k'}^4(X)}.$$

Since $E\eta_k^4(X)$ is uniformly bounded and $\sum_{k=s+1}^{\infty}\lambda_k^2 = o\left(\frac{1}{s}\right)$, we have

$$E\sum_{k=1}^{n}\left(\lambda_{nk} - \widetilde{\lambda}_{nk}^{(s)}\right)^2 \leq \frac{1}{n}\left(\sum_{k=s+1}^{\infty}\lambda_k\sqrt{E\eta_k^4(X)}\right)^2 + \frac{n-1}{n}\sum_{k=s+1}^{\infty}\lambda_k^2.$$

Hence, $\sum_{k=1}^{n}\left(\lambda_{nk} - \widetilde{\lambda}_{nk}^{(s)}\right)^2 \leq \operatorname{tr}\left(\widetilde{\boldsymbol{M}}_n^{(s)} - \boldsymbol{M}_n\right)^2 = O_p\left(\epsilon_{n,s}\right)$.

Next, for a fixed k,

$$\|\widetilde{\boldsymbol{M}}_n^{(s)}\boldsymbol{u}_{n,k} - \boldsymbol{M}_n\boldsymbol{u}_{n,k}\|_2^2 \leq \|\widetilde{\boldsymbol{M}}_n^{(s)} - \boldsymbol{M}_n\|_F^2 = O_p(\epsilon_{m,n}).$$

Thus, we have

$$\widetilde{\boldsymbol{M}}_n^{(s)}\boldsymbol{u}_{n,k} = \boldsymbol{M}_n\boldsymbol{u}_{n,k} + O_p^{L_2}\left(\sqrt{\epsilon_{m,n}}\right) = \lambda_{nk}\boldsymbol{u}_{n,k} + O_p^{L_2}\left(\sqrt{\epsilon_{m,n}}\right).$$

Now, let $\boldsymbol{u}_{nk} = \sum_{i=1}^{n}\alpha_i\widetilde{\boldsymbol{u}}_{ni}^{(s)}$, where $\sum_{i=1}^{n}\alpha_i^2 = 1$. Then, $\widetilde{\boldsymbol{M}}_n^{(s)}\boldsymbol{u}_{n,k} = \sum_{i=1}^{n}\alpha_i\,\widetilde{\lambda}_{ni}^{(s)}\,\widetilde{\boldsymbol{u}}_{ni}^{(s)}$. Therefore, we have

$$\sum_{i=1}^{n}\alpha_i(\lambda_{nk} - \widetilde{\lambda}_{ni}^{(s)})\widetilde{\boldsymbol{u}}_{ni}^{(s)} = \lambda_{nk}\boldsymbol{u}_{nk} - \widetilde{\boldsymbol{M}}_n^{(s)}\boldsymbol{u}_{n,k} = O_p^{L_2}\left(\sqrt{\widetilde{\epsilon}_{m,n}}\right),$$

which implies, by estimating its L_2 norm, that $\sum_{i=1}^{n}\alpha_i^2(\lambda_{nk} - \widetilde{\lambda}_{ni}^{(s)})^2 = O_p(\epsilon_{m,n})$. Note that, by Eqs. (19) and (24),[1] we have $(\lambda_k - \lambda_{nk})^2 = O_p(\epsilon_{m,n}) + O_p(sn^{-\nu})$ and therefore

$$\begin{aligned}\sum_{i=1}^{n}\alpha_i^2(\lambda_k - \lambda_i)^2 &\leq 3\sum_{i=1}^{n}\alpha_i^2(\lambda_k - \lambda_{nk})^2 + 3\sum_{i=1}^{n}\alpha_i^2(\lambda_{nk} - \widetilde{\lambda}_{ni}^{(s)})^2 + 3\sum_{i=1}^{n}(\widetilde{\lambda}_{ni}^{(s)} - \lambda_i)^2 \\ &= 3(\lambda_k - \lambda_{nk})^2 + O_p(\epsilon_{m,n}) + O_p(sn^{-\nu}) \\ &= O_p(\epsilon_{m,n}) + O_p(sn^{-\nu}).\end{aligned}$$

Then, for a fixed k, because of the distinctness assumption of λ_i's (see Eq. (4)), these α_i's must satisfy

$$\sum_{i\neq k}\alpha_i^2 = O_p(\epsilon_{m,n}) + O_p(sn^{-\nu}) \quad \text{and} \quad \alpha_k^2 = 1 + O_p(\epsilon_{m,n}) + O_p(sn^{-\nu}).$$

Without loss of generality, we can assume that $\alpha_k > 0$, so that $\alpha_k = 1 + O_p(\epsilon_{m,n}) + O_p(sn^{-\nu})$. (Note that each eigenvector has two directions so we can always choose the appropriate direction of $\boldsymbol{u}_{nk}$ to make $\alpha_k > 0$.) Thus, we have

$$\|\widetilde{\boldsymbol{u}}_{nk}^{(s)} - \boldsymbol{u}_{nk}\|_2^2 = (1-\alpha_k)^2 + \sum_{i\neq k}\alpha_i^2 = O_p(\epsilon_{m,n}) + O_p(sn^{-\nu}),$$

and we obtain Eq. (21). □

Lemma 3. *Let $s = n^{\alpha}$ with $0 < \alpha < \frac{\tau}{2(4+2\tau)}$. Then, for any $\alpha < \nu < \frac{\tau}{4+2\tau} - \alpha$, we have*

$$\sum_{k=1}^{s}\left(\lambda_k - \widetilde{\lambda}_{nk}^{(s)}\right)^2 = O_p\left(sn^{-\nu}\right). \tag{24}$$

[1] Although Eq. (24) is in Lemma 3, the proof of the equation does not require Lemma 2.

Furthermore, for a fixed k, we have

$$\widetilde{\lambda}_{nk}^{(s)} = \lambda_k + O_p\left(\sqrt{sn^{-2\nu}}\right), \tag{25}$$

$$\widetilde{\boldsymbol{u}}_{nk}^{(s)} = \frac{1}{\sqrt{n}}\eta_k(\boldsymbol{X}_n) + O_p^{L_2}\left(\sqrt{\epsilon_{n,s}}\right) + O_p^{L_2}\left(\sqrt{sn^{-2\nu}}\right). \tag{26}$$

Proof. From Lemma 1, when $n^{-(\nu-\alpha)}$ is small enough, we may assume that $\left(\boldsymbol{A}_n^{(s)}\right)^{1/2}$ is nonsingular. Thus, let $\boldsymbol{W}_n^{(s)} = \boldsymbol{V}_n^{(s)}\left(\boldsymbol{A}_n^{(s)}\right)^{-1/2}$ and then $\boldsymbol{W}_n^{(s)}$ is an $n \times s$ orthogonal matrix, i.e., $\boldsymbol{W}_n^{(s)T}\boldsymbol{W}_n^{(s)} = \boldsymbol{I}_s$. Note that $\widetilde{\boldsymbol{M}}_n^{(s)} = \boldsymbol{V}_n^{(s)}\boldsymbol{\Lambda}_s\boldsymbol{V}_n^{(s)T}$, so the kernel of $\widetilde{\boldsymbol{M}}_n^{(s)}$ contains $(\mathrm{span}\{\boldsymbol{V}_n^{(s)}\})^{\perp}$. Therefore, the first s eigenvalues of $\widetilde{\boldsymbol{M}}_n^{(s)}$ are the same as those of $\boldsymbol{W}_n^{(s)T}\widetilde{\boldsymbol{M}}_n^{(s)}\boldsymbol{W}_n^{(s)} = \left(\boldsymbol{A}_n^{(s)}\right)^{1/2}\boldsymbol{\Lambda}_s\left(\boldsymbol{A}_n^{(s)}\right)^{1/2}$, where $\boldsymbol{\Lambda}_s = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_s)$. By Lemma 2.3 of Bai [3],

$$\sum_{k=1}^{s}\left(\lambda_k - \widetilde{\lambda}_{nk}^{(s)}\right)^2 \le \mathrm{tr}\left(\boldsymbol{\Lambda}_s - \left(\boldsymbol{A}_n^{(s)}\right)^{1/2}\boldsymbol{\Lambda}_s\left(\boldsymbol{A}_n^{(s)}\right)^{1/2}\right)^2.$$

Denote the entries of $\boldsymbol{A}_n^{(s)}$ and $\left(\boldsymbol{A}_n^{(s)}\right)^{1/2}$ by a_{ij} and b_{ij}, respectively, i.e.,

$$\boldsymbol{A}_n^{(s)} := \left(a_{ij}\right)_{i,j=1}^{s}, \quad \text{and} \quad \left(\boldsymbol{A}_n^{(s)}\right)^{1/2} := \boldsymbol{B} := \left(b_{ij}\right)_{i,j=1}^{s}.$$

Therefore, we have, by simple calculation,

$$\begin{aligned}
\sum_{k=1}^{s}\left(\lambda_k - \widetilde{\lambda}_{nk}^{(s)}\right)^2 &\le \mathrm{tr}\left(\boldsymbol{\Lambda}_s - \boldsymbol{B}\boldsymbol{\Lambda}_s\boldsymbol{B}\right)^2 \\
&= \sum_{i=1}^{s}\lambda_i^2 - 2\sum_{i,k=1}^{s}\lambda_i\lambda_k b_{ik}^2 + \sum_{k,k'=1}^{s}\lambda_k\lambda_{k'}\underbrace{\left(\sum_{i=1}^{s}b_{ki}b_{k'i}\right)^2}_{a_{kk'}^2} \\
&= 2\sum_{i=1}^{s}\lambda_i^2 - 2\sum_{i,k=1}^{s}\lambda_i\lambda_k b_{ik}^2 + \sum_{k,k'=1}^{s}\lambda_k\lambda_{k'}a_{kk'}^2 - \sum_{i=1}^{s}\lambda_i^2 \\
&\le \left|-2\sum_{i,k=1}^{s}\lambda_i\lambda_k(b_{ki}^2 - \delta_{ki})\right| + \left|\sum_{i,k=1}^{s}\lambda_i\lambda_k(a_{ki}^2 - \delta_{ki})\right| \\
&\le \left(2\sup_{i,j}|b_{ij}^2 - \delta_{ij}| + \sup_{i,j}|a_{ij}^2 - \delta_{ij}|\right)\cdot\left(\sum_{i=1}^{\infty}\lambda_i\right)^2 \\
&= O_p\left(sn^{-\nu}\right).
\end{aligned}$$

Next, let us first consider the following expressions:

$$\begin{aligned}
\widetilde{\boldsymbol{M}}_n^{(s)}\cdot\frac{1}{\sqrt{n}}\eta_k(\boldsymbol{X}_n) &= \boldsymbol{V}_n^{(s)}\boldsymbol{\Lambda}_s\boldsymbol{V}_n^{(s)T}\cdot\frac{1}{\sqrt{n}}\eta_k(\boldsymbol{X}_n) \\
&= \boldsymbol{V}_n^{(s)}\begin{pmatrix}\lambda_1 a_{1k}\\ \lambda_2 a_{2k}\\ \vdots\\ \lambda_s a_{sk}\end{pmatrix} = \boldsymbol{V}_n^{(s)}\left[(0,\ldots,\lambda_k,\ldots,0)^T + \boldsymbol{\xi}\right],
\end{aligned}$$

where $\boldsymbol{\xi} = O_p^{(\infty)}\left(n^{-\nu}\right)$ by Lemma 1. Let $\boldsymbol{w} = \boldsymbol{V}_n^{(s)}\boldsymbol{\xi}$. Then, still by Lemma 1, we have

$$\begin{aligned}
\|\boldsymbol{w}\|_2^2 &= \left\langle\boldsymbol{V}_n^{(s)}\boldsymbol{\xi}, \boldsymbol{V}_n^{(s)}\boldsymbol{\xi}\right\rangle = \left\langle\boldsymbol{A}_n^{(s)}\boldsymbol{\xi}, \boldsymbol{\xi}\right\rangle \le \|\boldsymbol{\xi}\|_2^2 + \left|\boldsymbol{\xi}^T\left(\boldsymbol{A}_n^{(s)} - \boldsymbol{I}_m\right)\boldsymbol{\xi}\right| \\
&= O_p\left(\frac{s}{n^{2\nu}}\right) + O_p\left(\frac{sn^{-\nu}}{n^{2\nu}}\right) = O_p\left(\frac{s}{n^{2\nu}}\right).
\end{aligned}$$

Therefore,

$$\widetilde{\boldsymbol{M}}_n^{(s)}\cdot\frac{1}{\sqrt{n}}\eta_k(\boldsymbol{X}_n) = \lambda_k\cdot\frac{1}{\sqrt{n}}\eta_k(\boldsymbol{X}_n) + O_p^{(L_2)}\left(\sqrt{\frac{s}{n^{2\nu}}}\right). \tag{27}$$

Now let $\frac{1}{\sqrt{n}}\eta_k(\boldsymbol{X}_n) = \sum_{i=1}^n \alpha_i \widetilde{\boldsymbol{u}}_{ni}^{(s)}$. Then,

$$\widetilde{\boldsymbol{M}}_n^{(s)} \cdot \frac{1}{\sqrt{n}}\,\eta_k(\boldsymbol{X}_n) = \sum_{i=1}^n \alpha_i\, \widetilde{\lambda}_{ni}^{(s)}\, \widetilde{\boldsymbol{u}}_{ni}^{(s)}. \tag{28}$$

Due to Eq. (27), we have

$$\sum_{i=1}^n \alpha_i(\lambda_k - \widetilde{\lambda}_{ni}^{(s)})\widetilde{\boldsymbol{u}}_{ni}^{(s)} = \frac{1}{\sqrt{n}}\left(\lambda_k\eta_k(\boldsymbol{X}_n) - \widetilde{\boldsymbol{M}}_n^{(s)}\eta_k(\boldsymbol{X}_n)\right) = O_p^{(L_2)}\left(\sqrt{\frac{s}{n^{2\nu}}}\right),$$

which implies, by estimating its L_2 norm, that $\sum_{i=1}^n \alpha_i^2(\lambda_k - \widetilde{\lambda}_{ni}^{(s)})^2 = O_p\left(sn^{-2\nu}\right)$. Similar to the proof of Lemma 2, by Eqs. (19) and (24), we have

$$\begin{aligned}\sum_{i=1}^n \alpha_i^2(\lambda_k - \lambda_i)^2 &\le 3\sum_{i=1}^n \alpha_i^2(\lambda_k - \lambda_{nk})^2 + 3\sum_{i=1}^n \alpha_i^2(\lambda_{nk} - \widetilde{\lambda}_{ni}^{(s)})^2 + 3\sum_{i=1}^n (\widetilde{\lambda}_{ni}^{(s)} - \lambda_i)^2 \\ &= O_p(\epsilon_{m,n}) + O_p(sn^{-\nu}).\end{aligned}$$

Then, for a fixed k, because of the distinctness assumption of λ_i's (see Eq. (4)), these α_i's must satisfy

$$\begin{aligned}&\sum_{i\neq k} \alpha_i^2 = O_p(\epsilon_{n,s}) + O_p\left(sn^{-2\nu}\right), \\ &\alpha_k^2 = 1 + O_p(\epsilon_{n,s}) + O_p\left(sn^{-2\nu}\right).\end{aligned} \tag{29}$$

Without loss of generality, we can assume that $\alpha_k > 0$, by choosing an appropriate direction of $\widetilde{\boldsymbol{u}}_{nk}^{(s)}$, so $\alpha_k = 1 + O_p(\epsilon_{n,s}) + O_p\left(sn^{-2\nu}\right)$. Thus, we have

$$\left\|\widetilde{\boldsymbol{u}}_{nk}^{(s)} - \frac{1}{\sqrt{n}}\eta_k(\boldsymbol{X}_n)\right\|_2^2 = (1-\alpha_k)^2 + \sum_{i\neq k}\alpha_i^2 = O_p(\epsilon_{n,s}) + O_p\left(sn^{-2\nu}\right).$$

Hence, we obtain Eq. (26). □

Proof of Theorem 1. This theorem can be obtained by combining Lemmas 2 and 3. □

Next, recall the partition (6) and its rank-m approximation (7):

$$\boldsymbol{M}_n = \begin{bmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22}\end{bmatrix}, \qquad \widehat{\boldsymbol{M}}_n^{(m)} = \begin{pmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\end{pmatrix}.$$

Note that

$$\begin{aligned}\boldsymbol{M}_{11} &= \frac{1}{n}\sum_{k=1}^{\infty}\lambda_k\eta_k(\boldsymbol{X}_1^{(m)})\eta_k(\boldsymbol{X}_1^{(m)})^T, \\ \boldsymbol{M}_{12} &= \frac{1}{n}\sum_{k=1}^{\infty}\lambda_k\eta_k(\boldsymbol{X}_1^{(m)})\eta_k(\boldsymbol{X}_2^{(n\setminus m)})^T, \\ \boldsymbol{M}_{21} &= \frac{1}{n}\sum_{k=1}^{\infty}\lambda_k\eta_k(\boldsymbol{X}_2^{(n\setminus m)})\eta_k(\boldsymbol{X}_1^{(m)})^T, \\ \boldsymbol{M}_{22} &= \frac{1}{n}\sum_{k=1}^{\infty}\lambda_k\eta_k(\boldsymbol{X}_2^{(n\setminus m)}\boldsymbol{X}_2^{(n\setminus m)})\eta_k(\boldsymbol{X}_2^{(n\setminus m)}\boldsymbol{X}_2^{(n\setminus m)})^T,\end{aligned}$$

where $\eta_k(\boldsymbol{X}_1^{(m)}) := [\eta_k(X_1), \ldots, \eta_k(X_m)]^T$, and $\eta_k(\boldsymbol{X}_2^{(n\setminus m)}) := [\eta_k(X_{m+1}), \ldots, \eta_k(X_n)]^T$. Denote the m leading eigenvalues of $\widehat{\boldsymbol{M}}_n^{(m)}$ by $\widehat{\lambda}_{n1}^{(m)} \ge \cdots \ge \widehat{\lambda}_{nm}^{(m)}$ and the corresponding unit eigenvectors by $\widehat{\boldsymbol{u}}_{nk}^{(m)}$, $k = 1, 2, \ldots, m$.

Now, as expected, the term, $\boldsymbol{M}_{11}^{-1}$, in $\widehat{\boldsymbol{M}}_n^{(m)}$ is the key to make this approximation work, but it is also the most difficult term to be controlled because of its large eigenvalues. To break through this puzzle, the only place we can possibly connect with seems to be $\left\{\lambda_k, \frac{1}{n}\eta_k(\boldsymbol{X}_n)\right\}$. In Theorem 1, we have established the connection between $\{\lambda_{nk}, \boldsymbol{u}_{nk}\}$ and $\left\{\lambda_k, \frac{1}{n}\eta_k(\boldsymbol{X}_n)\right\}$. We will then use this connection to handle the eigenvalue problem of $\boldsymbol{M}_{11}^{-1}$. The key idea is to first truncate $\boldsymbol{M}_{12}$ and $\boldsymbol{M}_{21}$ so that we can have matrix decompositions (see (30)) and then to bring λ_i's of $\boldsymbol{M}_{12}$ and $\boldsymbol{M}_{21}$ appearing in $\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}$ to balance with the large eigenvalues of $\boldsymbol{M}_{11}^{-1}$.

Next, in order to use matrix operation, define the following two truncated matrices

$$\overline{\boldsymbol{M}}_n^{(m,l)} := \begin{pmatrix} \boldsymbol{M}_{11} & \widetilde{\boldsymbol{M}}_{12}^{(m,l)} \\ \widetilde{\boldsymbol{M}}_{21}^{(m,l)} & \widetilde{\boldsymbol{M}}_{22}^{(m,l)} \end{pmatrix}, \quad \text{and} \quad \widehat{\overline{\boldsymbol{M}}}_n^{(l)} := \begin{pmatrix} \boldsymbol{M}_{11} & \widetilde{\boldsymbol{M}}_{12}^{(m,l)} \\ \widetilde{\boldsymbol{M}}_{21}^{(m,l)} & \widetilde{\boldsymbol{M}}_{21}^{(m,l)} \boldsymbol{M}_{11}^{-1} \widetilde{\boldsymbol{M}}_{12}^{(m,l)} \end{pmatrix},$$

where

$$\widetilde{\boldsymbol{M}}_{12}^{(m,l)} := \frac{1}{n} \sum_{k=1}^{l} \lambda_k \eta_k(\boldsymbol{X}_1^{(m)}) \eta_k(\boldsymbol{X}_2^{(n\setminus m)})^T, \qquad \widetilde{\boldsymbol{M}}_{21}^{(m,l)} := \frac{1}{n} \sum_{k=1}^{l} \lambda_k \eta_k(\boldsymbol{X}_2^{(n\setminus m)}) \eta_k(\boldsymbol{X}_1^{(m)})^T,$$

$$\widetilde{\boldsymbol{M}}_{22}^{(m,l)} := \frac{1}{n} \sum_{k=1}^{l} \lambda_k \eta_k(\boldsymbol{X}_2^{(n\setminus m)}) \eta_k(\boldsymbol{X}_2^{(n\setminus m)})^T.$$

Notice that

$$\overline{\boldsymbol{M}}_n^{(m,l)} = \widetilde{\boldsymbol{M}}_n^{(l)} + \begin{pmatrix} \frac{1}{n} \sum_{k=l+1}^{\infty} \lambda_k \eta_k(\boldsymbol{X}_1^{(m)}) \eta_k(\boldsymbol{X}_1^{(m)})^T & 0_{m\times(n-m)} \\ 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)} \end{pmatrix},$$

where $0_{i\times j}$ is an $i \times j$ zero matrix. Since $\widetilde{\boldsymbol{M}}_n^{(l)}$ is nonnegative definite and

$$\frac{1}{n} \sum_{k=l+1}^{\infty} \lambda_k \eta_k(\boldsymbol{X}_1^{(m)}) \eta_k(\boldsymbol{X}_1^{(m)})^T$$

is positive definite, the matrix $\overline{\boldsymbol{M}}_n^{(m,l)}$ is nonnegative definite. Using the fact that if A is an m-square positive definite matrix, then

$$\begin{pmatrix} A & B^T \\ B & BA^{-1}B^T \end{pmatrix}$$

is nonnegative definite for any $(n-m) \times m$ matrix B (see Eq. (6.9) in [25]), the matrix $\widehat{\overline{\boldsymbol{M}}}_n^{(l)}$ is also nonnegative definite. Also note that $\widehat{\overline{\boldsymbol{M}}}_n^{(l)}$ has an interpretation as the difference of an unconditional covariance matrix minus a conditional covariance matrix, and hence, is nonnegative definite.[2]

Let $\boldsymbol{V}_n^{(l)}$ in (15) be partitioned into $\boldsymbol{V}_n^{(l)} = \begin{bmatrix} \boldsymbol{V}_1^{(l)} \\ \boldsymbol{V}_2^{(l)} \end{bmatrix}$, where

$$\boldsymbol{V}_1^{(l)} := \frac{1}{\sqrt{n}} \left(\eta_1(\boldsymbol{X}_1^{(m)}), \ldots, \eta_l(\boldsymbol{X}_1^{(m)}) \right)_{m\times l},$$

$$\boldsymbol{V}_2^{(l)} := \frac{1}{\sqrt{n}} \left(\eta_1(\boldsymbol{X}_2^{(n\setminus m)}), \ldots, \eta_l(\boldsymbol{X}_2^{(n\setminus m)}) \right)_{(n-m)\times l}.$$

Thus,

$$\widetilde{\boldsymbol{M}}_{12}^{(m,l)} = \boldsymbol{V}_1^{(l)} \boldsymbol{\Lambda}_l \boldsymbol{V}_2^{(l)T}, \qquad \widetilde{\boldsymbol{M}}_{21}^{(m,l)} = \boldsymbol{V}_2^{(l)} \boldsymbol{\Lambda}_l \boldsymbol{V}_1^{(l)T}, \qquad \widetilde{\boldsymbol{M}}_{22}^{(m,l)} = \boldsymbol{V}_2^{(l)} \boldsymbol{\Lambda}_l \boldsymbol{V}_2^{(l)T}. \tag{30}$$

Now, express $\boldsymbol{M}_{11}$ as follows:

$$\boldsymbol{M}_{11} = \frac{m}{n} \boldsymbol{U}_m \operatorname{diag}(\lambda_{m1}, \lambda_{m2}, \ldots, \lambda_{mm}) \boldsymbol{U}_m^T,$$

where $\boldsymbol{U}_m = (\boldsymbol{u}_{m1}, \boldsymbol{u}_{m2}, \ldots, \boldsymbol{u}_{m,m})$ consists of m orthogonal eigenvectors of $\frac{n}{m}\boldsymbol{M}_{11}$ with corresponding eigenvalues $\lambda_{m1} \geq \lambda_{m2} \geq \cdots \geq \lambda_{mm}$. Thus, $\boldsymbol{U}_m^T \boldsymbol{U}_m = \boldsymbol{I}_m$. Define $\boldsymbol{C}^{(l)}$ as follows: let

$$\begin{aligned} \widetilde{\boldsymbol{M}}_{21}^{(m,l)} \boldsymbol{M}_{11}^{-1} \widetilde{\boldsymbol{M}}_{12}^{(m,l)} &= \boldsymbol{V}_2^{(l)} \boldsymbol{\Lambda}_l \boldsymbol{V}_1^{(l)T} \boldsymbol{U}_m \frac{n}{m} \operatorname{diag}(\lambda_{m1}^{-1}, \lambda_{m2}^{-1}, \ldots, \lambda_{mm}^{-1}) \boldsymbol{U}_m^T \boldsymbol{V}_1^{(l)} \boldsymbol{\Lambda}_l \boldsymbol{V}_2^{(l)T} \\ &= \boldsymbol{V}_2^{(l)} \boldsymbol{C}^{(l)} \boldsymbol{V}_2^{(l)T}, \end{aligned}$$

---

[2] Suppose (X_1, X_2) has a multivariate normal distribution with covariance matrix $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. The covariance of X_2 conditional on X_1 is given by $\Sigma_{22\cdot 1} := \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

$$\operatorname{cov}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right) - \operatorname{cov}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \middle| X_1 \right) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_{22\cdot 1} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{bmatrix} \geq 0.$$

See, e.g., Anderson [2].

where $\boldsymbol{C}^{(l)}$ is an $l \times l$ matrix given by

$$\boldsymbol{C}^{(l)} = \frac{n}{m}\, \boldsymbol{\Lambda}_l \boldsymbol{V}_1^{(l)T} \boldsymbol{U}_m \operatorname{diag}(\lambda_{m1}^{-1}, \lambda_{m2}^{-1}, \ldots, \lambda_{mm}^{-1}) \boldsymbol{U}_m^T \boldsymbol{V}_1^{(l)} \boldsymbol{\Lambda}_l. \tag{31}$$

The (i, j)th entry of $\boldsymbol{C}^{(l)}$ is given by

$$c_{ij} = \lambda_i \lambda_j \sum_{k=1}^{m} \lambda_{mk}^{-1} q_{ki} q_{kj},$$

where q_{kj} is the (k, j)th entry of the matrix $\sqrt{\frac{n}{m}} \boldsymbol{U}_m^T \boldsymbol{V}_1^{(l)}$.

It is well known that $\widetilde{\boldsymbol{M}}_{22}^{(m,l)} \geq \widetilde{\boldsymbol{M}}_{21}^{(m,l)} \boldsymbol{M}_{11}^{-1} \widetilde{\boldsymbol{M}}_{12}^{(m,l)}$ for any m, n, l with $m < n$ (see Theorem 6.13 in [25]). Therefore, $\boldsymbol{V}_2^{(l)} (\boldsymbol{\Lambda}_l - \boldsymbol{C}^{(l)}) \boldsymbol{V}_2^{(l)T} \geq 0$ for any m, n, l with $m < n$. Now, by Lemma 1, $\boldsymbol{V}_2^{(l)T} \boldsymbol{V}_2^{(l)}$ is invertible infinitely often as n tends to infinity and $\boldsymbol{\Lambda}_l - \boldsymbol{C}^{(l)}$ is independent of $\boldsymbol{V}_2^{(l)}$, so we have $\boldsymbol{\Lambda}_l - \boldsymbol{C}^{(l)} \geq 0$. Thus,

$$\lambda_i = e_i^T \boldsymbol{\Lambda}_l e_i \geq e_i^T \boldsymbol{C}^{(l)} e_i = c_{ii}$$

for all i, where e_i is an l-vector with jth element $\delta_{i,j}$.

Lemma 4. *We have the following bounds for entries in $\boldsymbol{C}^{(l)}$, where $\boldsymbol{C}^{(l)}$ is defined in (31).*

$$c_{r_1 r_2} = \frac{1}{\sqrt{\lambda_{r_1} \lambda_{r_2}}} \left\{ O_p\left(\sqrt{\epsilon_{m,s_m}}\right) + O_p\left(\sqrt{s_m m^{-\nu}}\right) \right\}, \tag{32}$$

$$c_{rr} = \lambda_r + \frac{1}{\lambda_r} \left\{ O_p\left(\sqrt{\epsilon_{m,s_m}}\right) + O_p\left(\sqrt{s_m m^{-\nu}}\right) \right\}, \tag{33}$$

where $r_1 \neq r_2$.

Proof. We have the following inequality for c_{ij}:

$$c_{ij}^2 = \left(\sum_{k=1}^{m} \lambda_{mk}^{-1/2} \lambda_i q_{ki} \lambda_{mk}^{-1/2} \lambda_j q_{kj} \right)^2 \leq \sum_{k=1}^{m} \lambda_{mk}^{-1} \lambda_i^2 q_{ki}^2 \times \sum_{k=1}^{m} \lambda_{mk}^{-1} \lambda_j^2 q_{kj}^2 = c_{ii} c_{jj} \leq \lambda_i \lambda_j. \tag{34}$$

However, for any fixed integer $r > 0$,

$$\begin{aligned} \lambda_r \geq c_{rr} &= \lambda_r^2 \lambda_{mr}^{-1} q_{rr}^2 + \sum_{k \in \{1,2,\ldots,m\} \setminus \{r\}} \lambda_{mk}^{-1} \lambda_r^2 q_{kr}^2 \\ &\geq \lambda_r^2 \lambda_{mr}^{-1} q_{rr}^2 \\ &= \lambda_r + \frac{1}{\lambda_r} O_p\left(\sqrt{\epsilon_{m,s_m}}\right) + \frac{1}{\lambda_r} O_p\left(\sqrt{s_m m^{-\nu}}\right), \end{aligned} \tag{35}$$

where $s_m = m^{\alpha}$ with $0 < \alpha < \frac{\tau}{2(4+2\tau)}$ and $\alpha < \nu < \tau/(4+2\tau) - \alpha$. (Note that from the second line to third line, we use Theorem 1.) This implies that

$$c_{rr} = \lambda_r + \frac{1}{\lambda_r} \left\{ O_p\left(\sqrt{\epsilon_{m,s_m}}\right) + O_p\left(\sqrt{s_m m^{-\nu}}\right) \right\}$$

and

$$\sum_{k \in \{1,2,\ldots,m\} \setminus \{r\}} \lambda_{mk}^{-1} \lambda_r^2 q_{kr}^2 = \frac{1}{\lambda_r} \left\{ O_p\left(\sqrt{\epsilon_{m,s_m}}\right) + O_p\left(\sqrt{s_m m^{-\nu}}\right) \right\}.$$

Moreover, for fixed positive integers r_1, r_2,

$$\begin{aligned} c_{r_1 r_2} &= \lambda_{r_1} \lambda_{r_2} \sum_{k=1}^{m} \lambda_{mk}^{-1} q_{kr_1} q_{kr_2} \\ &= \lambda_{r_1} \lambda_{r_2} \lambda_{mr_1}^{-1} q_{r_1 r_1} q_{r_1 r_2} + \lambda_{r_1} \lambda_{r_2} \lambda_{mr_2}^{-1} q_{r_2 r_1} q_{r_2 r_2} + \lambda_{r_1} \lambda_{r_2} \sum_{k \in \{1,2,\ldots,m\} \setminus \{r_1, r_2\}} \lambda_{mk}^{-1} q_{kr_1} q_{kr_2}. \end{aligned}$$

Next, by replacing r by r_1 in Eq. (35), we can obtain

$$\lambda_{r_1} \lambda_{mr_1}^{-1} q_{r_1 r_1} \leq q_{r_1 r_1}^{-1},$$

so that the first term can be bounded as shown below:

$$\lambda_{r_1}\lambda_{r_2}\lambda_{mr_1}^{-1}q_{r_1r_1}q_{r_1r_2} \le \lambda_{r_2}q_{r_1r_2}q_{r_1r_1}^{-1} = O_p\left(\sqrt{\epsilon_{m,s_m}}\right) + O_p\left(\sqrt{s_m m^{-\nu}}\right).$$

Similarly, the second term

$$\lambda_{r_2}\lambda_{r_1}\lambda_{mr_2}^{-1}q_{r_2r_2}q_{r_2r_1} = O_p\left(\sqrt{\epsilon_{m,s_m}}\right) + O_p\left(\sqrt{s_m m^{-\nu}}\right).$$

Now for the third term, we can again use Cauchy–Schwarz inequality as (35) and obtain

$$\lambda_{r_1}\lambda_{r_2}\sum_{k\in\{1,2,\ldots,m\}\setminus\{r_1,r_2\}}\lambda_{mk}^{-1}q_{kr_1}q_{kr_2} = \frac{1}{\sqrt{\lambda_{r_1}\lambda_{r_2}}}\left\{O_p\left(\epsilon_{m,s_m}\right) + O_p\left(s_m m^{-\nu}\right)\right\}.$$

Hence, we have the following bound for fixed positive integers r_1 and r_2:

$$c_{r_1r_2} = \frac{1}{\sqrt{\lambda_{r_1}\lambda_{r_2}}}\left\{O_p\left(\sqrt{\epsilon_{m,s_m}}\right) + O_p\left(\sqrt{s_m m^{-\nu}}\right)\right\}. \quad \square$$

Proof of Theorem 2.

$$\begin{aligned}\sum_{k=1}^{n}\left(\lambda_{nk} - \widehat{\lambda}_{nk}^{(m)}\right)^2 &\le \operatorname{tr}\left(\boldsymbol{M}_n - \widehat{\boldsymbol{M}}_n^{(m)}\right)^2 = \|\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\|_F^2\\ &= \frac{1}{n^2}\sum_{i,j=1}^{n-m}\sum_{k_1,k_1',k_2,k_2'=1}^{\infty}(\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})(\lambda_{k_1'}\delta_{k_1',k_2'} - c_{k_1',k_2'})\\ &\quad\times \eta_{k_1}(X_{m+i})\eta_{k_1'}(X_{m+i})\eta_{k_2}(X_{m+j})\eta_{k_2'}(X_{m+j}).\end{aligned}$$

Let $G = \eta_{k_1}(X_{m+i})\eta_{k_1'}(X_{m+i})\eta_{k_2}(X_{m+j})\eta_{k_2'}(X_{m+j})$. Since $\sup_k E\eta_k^4(X)$ is bounded by M, we can obtain $|E(G)| \le M$ if $i = j$. Furthermore,

$$E(G) = \begin{cases}1, & \text{if } (k_1,k_2) = (k_1',k_2'),\ i \le j,\\ 0, & \text{if } (k_1,k_2) \ne (k_1',k_2'),\ i \ne j.\end{cases} \tag{36}$$

Therefore,

$$\begin{aligned}E\left(\|\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\|_F^2\big|\boldsymbol{X}_1^m\right) &= \frac{1}{n^2}\sum_{i,j=1}^{n-m}\sum_{k_1,k_1',k_2,k_2'=1}^{\infty}(\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})(\lambda_{k_1'}\delta_{k_1',k_2'} - c_{k_1',k_2'})E(G)\\ &\le I_1 + I_2,\end{aligned}$$

where

$$I_1 = \left(\frac{n-m}{n}\right)^2\sum_{k_1,k_2=1}^{\infty}(\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})^2 \quad (\text{for } i \ne j)$$

and

$$I_2 = \frac{n-m}{n^2}M\left(\sum_{k_1,k_2=1}^{\infty}(\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})\right)^2 \quad (\text{for } i = j).$$

Now we write $I_1 = I_{11} + I_{12} + I_{13}$, where

$$I_{11} = \left(\frac{n-m}{n}\right)^2\sum_{k_1,k_2=1}^{l}(\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})^2,$$

$$I_{12} = \left(\frac{n-m}{n}\right)^2\sum_{k_1,k_2=l+1}^{\infty}(\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})^2,$$

$$I_{13} = 2\left(\frac{n-m}{n}\right)^2\sum_{k_1=l+1}^{\infty}\sum_{k_2=1}^{l}(\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})^2.$$

By Eq. (32) in Lemma 4, the first term

$$I_{11} = \left(\frac{n-m}{n}\right)^2 \left(\sum_{r=1}^{l} \frac{1}{\lambda_r}\right)^2 \left\{O_p\left(\epsilon_{m,s_m}\right) + O_p\left(s_m m^{-\nu}\right)\right\}. \tag{37}$$

For the second term, since

$$\sum_{k_1,k_2=l+1}^{\infty} (\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})^2 \le 4\left(\sum_{k_1,k_2=l+1}^{\infty} \lambda_{k_1}^2\delta_{k_1,k_2} + \sum_{k_1,k_2=l+1}^{\infty} c_{k_1k_2}^2\right)$$

and $c_{ij}^2 \le \lambda_i\lambda_j$,

$$I_{12} \le \left(\frac{n-m}{n}\right)^2 \left[\sum_{k=l+1}^{\infty} \lambda_k^2 + \left(\sum_{k=l+1}^{\infty} \lambda_k\right)^2\right]. \tag{38}$$

For the third term,

$$\begin{aligned} I_{13} &= 2\left(\frac{n-m}{n}\right)^2 \sum_{k_1=l+1}^{\infty}\sum_{k_2=1}^{l} c_{k_1k_2}^2 \le 2\left(\frac{n-m}{n}\right)^2 \sum_{k_1=l+1}^{\infty}\sum_{k_2=1}^{l} \lambda_{k_1}\lambda_{k_2} \\ &\le 2\left(\frac{n-m}{n}\right)^2 \left(\sum_{k_2=1}^{\infty} \lambda_{k_2}\right)\left(\sum_{k_1=l+1}^{\infty} \lambda_{k_1}\right). \end{aligned} \tag{39}$$

Therefore, by Eqs. (37)–(39), we have

$$I_1 = \left(\frac{n-m}{n}\right)^2 \left[\left(\sum_{r=1}^{l} \frac{1}{\lambda_r}\right)^2 \left(O_p\left(\epsilon_{m,s_m}\right) + O_p\left(s_m m^{-\nu}\right)\right) + O\left(\sum_{k=l+1}^{\infty} \lambda_k\right)\right]. \tag{40}$$

Now let us consider I_2.

$$I_2 = \frac{n-m}{n^2} M\left(\lim_{l\to\infty} \sum_{k_1,k_2=1}^{l} (\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2})\right)^2.$$

For any integer $l > 0$, the finite sum

$$\sum_{k_1,k_2=1}^{l} (\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2}) = (1, 1, \ldots, 1)(\boldsymbol{\Lambda}_l - \boldsymbol{C}^{(l)})(1, 1, \ldots, 1)^T.$$

Since $\boldsymbol{\Lambda}_l \ge \boldsymbol{C}^{(l)}$, we have

$$0 \le \sum_{k_1,k_2=1}^{l} (\lambda_{k_1}\delta_{k_1,k_2} - c_{k_1k_2}) \le 2(1, 1, \ldots, 1)\boldsymbol{\Lambda}_l(1, 1, \ldots, 1)^T \le 2\sum_{i=1}^{\infty} \lambda_i.$$

Thus,

$$I_2 \le 2M\frac{n-m}{n^2}\sum_{i=1}^{\infty} \lambda_i = O\left(\frac{n-m}{n^2}\right). \tag{41}$$

Together with Eqs. (40) and (41), we obtain

$$\begin{aligned} E(\|\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\|_F^2 | X_1^m) = \left(\frac{n-m}{n}\right)^2 \Bigg[\left(\sum_{r=1}^{l} \frac{1}{\lambda_r}\right)^2 \left(O_p\left(\epsilon_{m,s_m}\right) + O_p\left(s_m m^{-\nu}\right)\right) \\ + O\left(\sum_{k=l+1}^{\infty} \lambda_k\right)\Bigg] + O\left(\frac{n-m}{n^2}\right). \end{aligned}$$

Note that, we can choose $l = l_m$ as a function of m which tends to infinity slow enough as $m \to \infty$ such that

$$\lim_{m\to\infty}\left[\left(\sum_{r=1}^{l_m}\frac{1}{\lambda_r}\right)^2\left(\epsilon_{m,s_m}+s_m m^{-\nu}\right)+\sum_{k=l_m+1}^{\infty}\lambda_k\right]=0.$$

Hence, we obtain

$$\sum_{i=1}^{n}\left(\lambda_{ni}-\widehat{\lambda}_{ni}^{(m)}\right)^2\le\|\boldsymbol{M}_n-\widehat{\boldsymbol{M}}_n^{(m)}\|_F^2=O_p(\tilde{\epsilon}_{m,n}).$$

This gives us Eqs. (12) and (13).

Next, let us first consider, for a fixed k,

$$\|\widehat{\boldsymbol{M}}_n^{(m)}\boldsymbol{u}_{nk}-\boldsymbol{M}_n\boldsymbol{u}_{nk}\|_2^2\le\|\widehat{\boldsymbol{M}}_n^{(m)}-\boldsymbol{M}_n\|_F^2=O_p(\tilde{\epsilon}_{m,n}).$$

Thus, we have

$$\widehat{\boldsymbol{M}}_n^{(m)}\boldsymbol{u}_{nk}=\boldsymbol{M}_n\boldsymbol{u}_{nk}+O_p^{L_2}\left(\sqrt{\tilde{\epsilon}_{m,n}}\right)=\lambda_{nk}\boldsymbol{u}_{nk}+O_p^{L_2}\left(\sqrt{\tilde{\nu}_{m,n}}\right).$$

Similar to the deviation of Eq. (10), we let $\boldsymbol{u}_{nk}=\sum_{i=1}^n\alpha_i\widehat{\boldsymbol{u}}_{ni}$. Then, $\widehat{\boldsymbol{M}}_n^{(m)}\boldsymbol{u}_{nk}=\sum_{i=1}^n\alpha_i\,\widehat{\lambda}_{ni}^{(m)}\,\widehat{\boldsymbol{u}}_{ni}$. Therefore, we have

$$\sum_{i=1}^{n}\alpha_i(\lambda_{nk}-\widehat{\lambda}_{ni}^{(m)})\widehat{\boldsymbol{u}}_{ni}=\lambda_{nk}\boldsymbol{u}_{nk}-\widehat{\boldsymbol{M}}_n^{(m)}\boldsymbol{u}_{nk}=O_p^{L_2}\left(\sqrt{\tilde{\epsilon}_{m,n}}\right),$$

which implies, by estimating its L_2 norm, that $\sum_{i=1}^n\alpha_i^2(\lambda_k-\lambda_{ni})^2=O_p(\tilde{\epsilon}_{m,n})$. Note that, by Theorem 1 and Eq. (13), we have

$$\lambda_{nk}-\widehat{\lambda}_{ni}^{(m)}=\lambda_k-\lambda_i+O_p\left(\sqrt{\epsilon_{n,s}}\right)+O_p\left(\sqrt{sn^{-\nu}}\right)+O_p\left(\sqrt{\tilde{\epsilon}_{m,n}}\right).$$

Then, for a fixed k, these α_i's must satisfy

$$\sum_{i\neq k}\alpha_i^2=O_p(\tilde{\epsilon}_{m,n})\quad\text{and}\quad\alpha_k^2=1+O_p(\tilde{\epsilon}_{m,n}).$$

Without loss the generality, we can again assume that $\alpha_k > 0$, by choosing an appropriate direction of $\widehat{\boldsymbol{u}}_{nk}$, so that $\alpha_k=1+O_p(\tilde{\epsilon}_{m,n})$. Thus, we have

$$\|\widehat{\boldsymbol{u}}_{nk}-\boldsymbol{u}_{nk}\|_2^2=(1-\alpha_k)^2+\sum_{i\neq k}\alpha_i^2=O_p(\tilde{\epsilon}_{m,n}),$$

and we obtain Eq. (14). □

4. Examples and numerical study

Example 1. Consider the covariance kernel of the Brownian motion $K(t,s) := t\wedge s$, for $t,s\in[0,1]$ under uniform distribution $U(0,1)$. By the Fourier theory, $\left\{\sqrt{2}\sin\left[\left(k+\frac{1}{2}\right)\pi t\right]\right\}_{k=0}^{\infty}$ forms a complete and orthonormal basis for the family of $L_2([0,1])$ functions. Thus, the indicator function

$$\mathbb{1}_{(0,t)}(u)=\sum_{k=0}^{\infty}\frac{4}{(2k+1)\pi}\cos\left[\left(k+\frac{1}{2}\right)\pi u\right]\cdot\sin\left[\left(k+\frac{1}{2}\right)\pi t\right],$$

which converges in the sense of $L_2([0,1])$. Thus, we have the following spectral decomposition of the kernel $K(s,t)$:

$$t\wedge s=\int_0^1\mathbb{1}_{(0,t)}(u)\mathbb{1}_{(0,s)}(u)du=\sum_{k=0}^{\infty}\frac{8}{(2k+1)^2\pi^2}\sin\left[\left(k+\frac{1}{2}\right)\pi t\right]\cdot\sin\left[\left(k+\frac{1}{2}\right)\pi s\right].$$

This equation shows that the eigenvalues for the kernel function, $K(t,s)=t\wedge s$, are $\lambda_k=\frac{4}{(2k+1)^2\pi^2}$ and the corresponding eigenfunctions are $\eta_k(t)=\sqrt{2}\sin\left[\left(k+\frac{1}{2}\right)\pi t\right]$. In Theorem 2, we have obtained that the error of Nyström approximation goes to zero as m, n tend to infinity through an upper bound with rate of $O(\tilde{\epsilon}_{m,n})$. However, $\tilde{\epsilon}_{m,n}$ does not reflect the actual rate of convergence as seen in the numerical experiment with Fig. 4. In this figure, we compute, for each (m,n), the absolute error of the largest eigenvalue and the L_2 error (Euclidean distance) of the corresponding unit eigenvector. The middle panel of the figure shows the rapid convergence of the largest eigenvalue and it corresponding eigenvector for the case of $m=\sqrt{n}$. By log-linear regression estimate (linear regression estimate for logarithmic errors), the error rate is about of $O(n^{-0.8})$, but

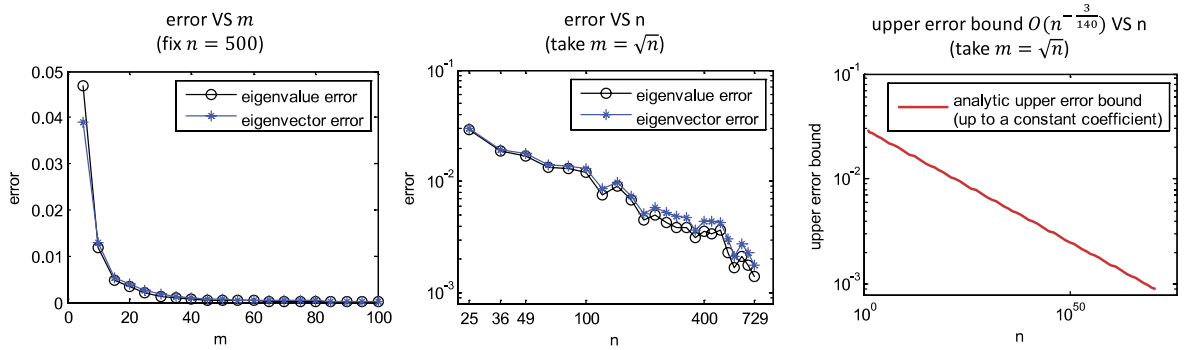

Fig. 4. Kernel K(t, s) := t ∧ s. The Nyström approximation errors of largest eigenvalues and their corresponding eigenvectors are plotted, respectively, by black circle and blue star. The left panel shows the fast decay of the errors as m increases toward the fixed $n = 500$. The middle panel is the log–log plot of error versus n and shows the errors tend to zeros as n, m tend to infinity with $m = \sqrt{n}$. The right panel is the log–log plot and shows the slow decay of the analytic upper error bound.

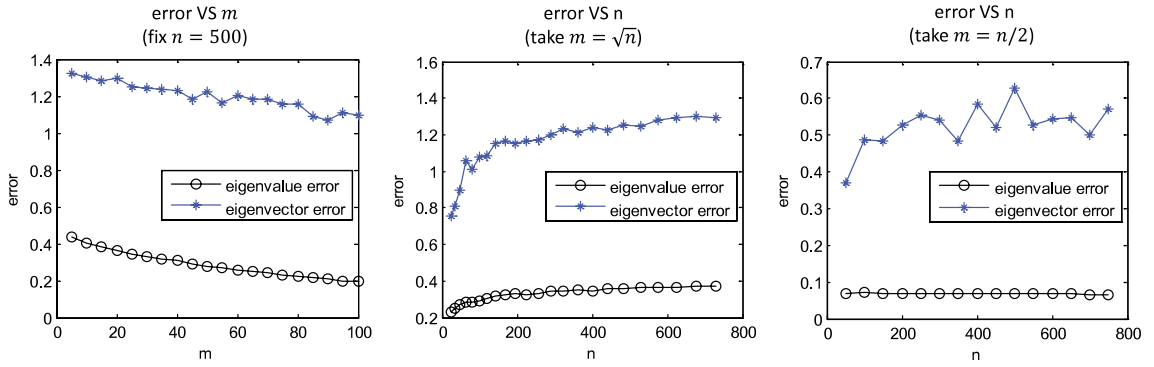

Fig. 5. Wishart matrix. The Nyström approximation errors of largest eigenvalues and their corresponding eigenvectors are plotted by black circle and blue star. The left panel shows the slow decay of the errors as m increases toward the fixed $n = 500$. The middle and right panel show the errors do not converge to zero when m, n tend to infinity with $m = \sqrt{n}$ and $m = n/2$.

by the formula of Remark 3, the upper bound rate $\tilde{\epsilon}_{m,n} = O\left(m^{-\frac{3}{70}(1-2\delta)}\right) = O\left(n^{-\frac{3}{140}(1-2\delta)}\right)$ which is much slower than the numerical experiment as we can see in the right panel of the figure. (Notice that in this example, $\beta = 2$ [i.e., $\lambda_k = O(k^{-2})$], τ can be arbitrarily large and δ can be arbitrarily small.) The left panel shows that given $n = 500$ fixed, the errors evanesce as m is greater than 20. Thus, taking $m = \sqrt{n}$ might be quite appropriate for this example.

Example 2. Not all symmetric, positive definite, random matrices can have a set of fixed underlying spectrums as the basic assumption of this paper. Let us consider the Wishart random matrix (scaled by $\frac{1}{n}$) as follows:

$$\boldsymbol{M}_n = \frac{1}{n}\boldsymbol{X}_n\boldsymbol{X}_n^T,$$

where $\boldsymbol{X}_n$ is an $n \times n$ matrix with i.i.d. standard normal random entries. When n tends to infinity, by Marchenko–Pastur law, the empirical distribution of the eigenvalues of $\boldsymbol{M}_n$ becomes dense in an interval. We can see in Fig. 5 that the Nyström method does not work, since these Wishart matrices are not derived, as in Eq. (5), from an underlying continuous kernel which has a discrete spectral decomposition (Eq. (2)).

Notation

- $\boldsymbol{A}_n^{(s)} = \left[a_{ij}\right]_{i,j=1}^{s} := \boldsymbol{V}_n^{(s)T}\boldsymbol{V}_n^{(s)}$.
- $\boldsymbol{C}^{(l)} := \boldsymbol{\Lambda}_l\boldsymbol{V}_1^{(l)T}\boldsymbol{M}_{11}^{-1}\boldsymbol{V}_1^{(l)}\boldsymbol{\Lambda}_l \in \Re^{l\times l}$, where $\boldsymbol{V}_1^{(l)} := \frac{1}{\sqrt{n}}\left[\eta_1(\boldsymbol{X}_1^{(m)}), \ldots, \eta_l(\boldsymbol{X}_1^{(m)}), \ldots\right]_{m\times l}$.
- $\eta_k(\boldsymbol{X}_1^{(m)}) := [\eta_k(X_1), \ldots, \eta_k(X_m)]^T$, and $\eta_k(\boldsymbol{X}_2^{(n\setminus m)}) := [\eta_k(X_{m+1}), \ldots, \eta_k(X_n)]^T$.
- $\boldsymbol{I}_s$ is the identity matrix with size $s \times s$.
- $\widetilde{K}^{(s)}(x, y) := \sum_{k=1}^{s}\lambda_k\eta_k(x)\eta_k(y)$, the truncated kernel at s components.
  $\widetilde{\boldsymbol{K}}_n^{(s)} := \left[\widetilde{K}^{(s)}(X_i, X_j)\right]_{i,j=1}^{n}$.

- λ_k, $\eta_k(x)$: the kth eigenvalue and associated eigenfunction of $K(x, y)$.
- λ_{nk}, $\boldsymbol{u}_{nk}$: the kth eigenvalue and associated eigenvector of $\boldsymbol{M}_n$.
- $\widehat{\lambda}_{nk}^{(m)}$, $\widehat{\boldsymbol{u}}_{nk}^{(m)}$: the kth eigenvalue and associated eigenvector of $\widehat{\boldsymbol{M}}_n^{(m)}$.
- $\widetilde{\lambda}_{nk}^{(s)}$, $\widetilde{\boldsymbol{u}}_{nk}^{(s)}$: the kth eigenvalue and associated eigenvector of $\widetilde{\boldsymbol{M}}_n^{(s)}$.
- $\boldsymbol{\Lambda}_s := \mathrm{diag}(\lambda_1, \ldots, \lambda_s) \in \Re^{s\times s}$.
- $\boldsymbol{M}_n := \frac{1}{n}K(\boldsymbol{X}_n, \boldsymbol{X}_n) = \frac{1}{n}\left[K(X_i, X_j)\right]_{i,j=1}^n = \frac{1}{n}\boldsymbol{K}_n$. Sometimes $\boldsymbol{M}_n$ is further partitioned into $\boldsymbol{M}_n = \begin{bmatrix}\boldsymbol{M}_{11} & \boldsymbol{M}_{12}\\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22}\end{bmatrix}$, where $\boldsymbol{M}_{11} \in \Re^{m\times m}$, $\boldsymbol{M}_{12} \in \Re^{m\times(n-m)}$, and $\boldsymbol{M}_{22} \in \Re^{(n-m)\times(n-m)}$.
- $\widehat{\boldsymbol{M}}_n^{(m)} := \begin{bmatrix}\boldsymbol{M}_{11}\\ \boldsymbol{M}_{21}\end{bmatrix}\boldsymbol{M}_{11}^{-1}\left[\boldsymbol{M}_{11}\ \boldsymbol{M}_{12}\right]$, which is a Nyström rank-m approximation to $\boldsymbol{M}_n$.
- $\widetilde{\boldsymbol{M}}_n^{(s)} := \frac{1}{n}\left[\widetilde{K}^{(s)}(X_i, X_j)\right]_{i,j=1}^n = \frac{1}{n}\widetilde{\boldsymbol{K}}_n^{(s)}$, an $n \times n$ matrix with rank at most s.
- $\overline{\boldsymbol{M}}_n^{(m,s)} := \begin{pmatrix}\boldsymbol{M}_{11} & \widetilde{\boldsymbol{M}}_{12}^{(m,s)}\\ \widetilde{\boldsymbol{M}}_{21}^{(m,s)} & \widetilde{\boldsymbol{M}}_{22}^{(m,s)}\end{pmatrix}$
- $\widetilde{\boldsymbol{M}}_{12}^{(m,s)} := \frac{1}{n}\sum_{k=1}^s \lambda_k\eta_k(\boldsymbol{X}_1^{(m)})\eta_k(\boldsymbol{X}_2^{(n\setminus m)})^T$, $\widetilde{\boldsymbol{M}}_{21}^{(m,s)} = \widetilde{\boldsymbol{M}}_{12}^{(m,s)T}$,
  $\widetilde{\boldsymbol{M}}_{22}^{(m,s)} := \frac{1}{n}\sum_{k=1}^s \lambda_k\eta_k(\boldsymbol{X}_2^{(n\setminus m)})\eta_k(\boldsymbol{X}_2^{(n\setminus m)})^T$.
- $\widehat{\overline{\boldsymbol{M}}}_n^{(s)} := \begin{pmatrix}\boldsymbol{M}_{11} & \widetilde{\boldsymbol{M}}_{12}^{(m,s)}\\ \widetilde{\boldsymbol{M}}_{21}^{(m,s)} & \widetilde{\boldsymbol{M}}_{21}^{(m,s)}\boldsymbol{M}_{11}^{-1}\widetilde{\boldsymbol{M}}_{12}^{(m,s)}\end{pmatrix}$.
- $\boldsymbol{V}_n^{(s)} := \frac{1}{\sqrt{n}}\left(\eta_1(\boldsymbol{X_n}), \eta_2(\boldsymbol{X_n}), \ldots, \eta_s(\boldsymbol{X_n})\right)$, which is an $n \times s$ matrix consisting of leading s eigenfunctions evaluated at data points $\boldsymbol{X}_n$ and scaled by $1/\sqrt{n}$. Sometimes $\boldsymbol{V}_n^{(s)}$ is further partitioned into 2 sub-matrices: $\boldsymbol{V}_n^{(s)} = \begin{bmatrix}\boldsymbol{V}_1^{(s)}\\ \boldsymbol{V}_2^{(s)}\end{bmatrix}$, where $\boldsymbol{V}_1^{(s)} \in \Re^{m\times s}$ and $\boldsymbol{V}_2^{(s)} \in \Re^{(n-m)\times s}$.
- $\boldsymbol{X}_n := [X_1, \ldots, X_n]^T$, which is the data design matrix. Sometimes this data design matrix is partitioned into two sub-matrices $\boldsymbol{X}_n = \begin{bmatrix}\boldsymbol{X}_1^{(m)}\\ \boldsymbol{X}_2^{(n\setminus m)}\end{bmatrix}_{n\times p}$, where $\boldsymbol{X}_1^{(m)} := [X_1, \ldots, X_m]^T \in \Re^{m\times p}$ and $\boldsymbol{X}_2^{(n\setminus m)} := [X_{m+1}, \ldots, X_n]^T \in \Re^{(n-m)\times p}$.
- $\widehat{\boldsymbol{X}}_n^{(m)}$: a data subset matrix of size $m \times p$ formed by a random subset of size m from $\{X_1, \ldots, X_n\}$. Since $X_1, \ldots, X_n$ are i.i.d. copies from X, without loss of generality, we may assume that $\widehat{\boldsymbol{X}}_n^{(m)} = \boldsymbol{X}_1^{(m)}$.

Acknowledgments

The author Lo-Bin Chang gratefully acknowledges the support from the Center of Mathematical Modeling & Scientific Computing, and the National Center for Theoretical Science, Hsinchu, Taiwan, and the National Science Council under grant 100-2115-M-009-007-MY2, and the Defense Advanced Research Projects Agency under contract FA8650-11-1-7151. The authors Chii-Ruey Hwang and Su-Yun Huang gratefully acknowledge the support from the National Science Council under grant 101-2115-M-001-012-MY2, 98-2115-M-001-007-MY3, 99-2118-M-001-007-MY2 and 101-2118-M-001-008.

References

[1] E. Alpaydm, Introduction to Machine Learning, MIT Press, Cambridge, MA, 2004.
[2] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, 2003.
[3] Z.D. Bai, Methodologies in spectral analysis of large dimensional random matrices, a review, Statistica Sinica 9 (1999) 611–677.
[4] C. Cortes, V.N. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[5] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000.
[6] P. Drineas, R. Kannan, M.W. Mahoney, Fast Monte Carlo algorithms for matrices I: approximating matrix multiplication, SIAM Journal on Computing 36 (2006) 132–157.
[7] P. Drineas, R. Kannan, M.W. Mahoney, Fast Monte Carlo algorithms for matrices II: computing a low-rank approximation to a matrix, SIAM Journal on Computing 36 (2006) 158–183.
[8] P. Drineas, R. Kannan, M.W. Mahoney, Fast Monte Carlo algorithms for matrices III: computing a compressed approximate matrix decomposition, SIAM Journal on Computing 36 (2006) 184–206.
[9] P. Drineas, M.W. Mahoney, On the Nyström method for approximating a Gram matrix for improved kernel-based learning, Journal of Machine Learning Research 6 (2005) 2153–2175.
[10] K.T. Fang, D.K.J. Lin, P. Winker, Y. Zhang, Uniform design: theory and application, Technometrics 42 (2000) 237–248.
[11] G. Fung, O.L. Mangasarian, Proximal support vector machine classifiers, in: KDD '01 Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, pp. 77–86.
[12] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, New York, 2009.
[13] S.Y. Huang, Y.R. Yeh, S. Eguchi, Robust kernel principal component analysis, Neural Computation 21 (2009) 3179–3213.
[14] Y.J. Lee, W.F. Hsieh, C.M. Huang, ϵ-SSVR: a smooth support vector machine for ϵ-insensitive regression, IEEE Transactions on Knowledge and Data Engineering 17 (2005) 678–685.
[15] Y.J. Lee, S.Y. Huang, Reduced support vector machines: a statistical theory, IEEE Transactions on Neural Networks 18 (2007) 1–13.
[16] Y.J. Lee, O.L. Mangasarian, RSVM: reduced support vector machines, in: Proceedings of the First SIAM International Conference on Data Mining, SIAM, Philadelphia, 2001.
[17] M.W. Mahoney, P. Drineas, CUR matrix decompositions for improved data analysis, Proceedings of the National Academy of Sciences of the United States of America 106 (3) (2009) 697–702.
[18] O.L. Mangasarian, D.R. Musicant, Lagrangian support vector machines, Journal of Machine Learning Research 1 (2001) 161–177.
[19] D.R. Musicant, A. Feinberg, Active set support vector regression, IEEE Transactions on Neural Networks 15 (2004) 268–275.

[20] B. Schölkopf, A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, 2002.
[21] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters 9 (1999) 293–300.
[22] J.A.K. Suykens, J. Vandewalle, Multiclass least squares support vector machines, in: Proc. IJCNN, Washington, DC, 1999, pp. 900–903.
[23] C. Williams, M. Seeger, Using the Nyström method to speed up kernel machines, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems, Vol. 13, MIT Press, Cambridge, MA, 2001, pp. 682–688.
[24] Y.R. Yeh, S.Y. Huang, Y.J. Lee, Nonlinear dimension reduction with kernel sliced inverse regression, IEEE Transactions on Knowledge and Data Engineering 21 (2009) 1590–1603.
[25] F. Zhang, Matrix Theory: Basic Results and Techniques, Springer, New York, 1999.