

Chapter 2

Matrix Analysis Basics

In this chapter, we review some basic concepts, properties, and theorems of singular value decomposition (SVD), eigenvalue decomposition (ED), and Rayleigh quotient of a matrix. Moreover, we also introduce some basics of matrix analysis. They are important and useful for our theoretical analysis in subsequent chapters.

2.1 Introduction

As discussed in Chap. 1, the PC or MC can be obtained by the ED of the sample correlation matrix or the SVD of the data matrix, and ED and SVD are also primal analysis tools. The history of SVD can date back to the 1870s, and Beltrami and Jordan are acknowledged as the founder of SVD. In 1873, Beltrami [1] published the first paper on SVD, and one year later Jordan [2] published his independent reasoning about SVD. Now, SVD has become one of the most useful and most efficient modern numerical analysis tools, and it has been widely used in statistical analysis, signal and image processing, system theory and control, etc. SVD is also a fundamental tool for eigenvector extraction, subspace tracking, and total least squares problem, etc.

On the other hand, ED is important in both mathematical analysis and engineering applications. For example, in matrix algebra, ED is usually related to the spectral analysis, and the spectral of a linear arithmetic operator is defined as the set of eigenvalues of the matrix. In engineering applications, spectral analysis is connected to the Fourier analysis, and the frequency spectral of signals is defined as the Fourier spectral, and then the power spectral of signals is defined as the square of frequency spectral norm or Fourier transform of the autocorrelation functions.

Besides SVD and ED, gradient and matrix differential are also the important concepts of matrix analysis. In view of the use of them in latter chapters, we will provide detailed analysis of SVD, ED, matrix analysis, etc. in the following.

2.2 Singular Value Decomposition

As to the inventor history of SVD, see Stewart's dissertation. Later, Autonne [3] extended SVD to complex square matrix in 1902, and Eckart and Young [4] further extended it to general rectangle matrix in 1939. Now, the theorem of SVD for rectangle matrix is usually called Eckart–Young Theorem.

SVD can be viewed as the extension of ED to the case of nonsquare matrices. It says that any real matrix can be diagonalized by using two orthogonal matrices. ED works only for square matrices and uses only one matrix (and its inverse) to achieve diagonalization. If the matrix is square and symmetric, then the two orthogonal matrices of SVD will be the same, and ED and SVD will also be the same and closely related to the matrix rank and reduced-rank least squares approximations.

2.2.1 Theorem and Uniqueness of SVD

Theorem 2.1 *For any $\mathbf{A} \in \mathbb{R}^{m \times n}$ (or $\mathbb{C}^{m \times n}$), there exist two orthonormal (or unitary) matrices $\mathbf{U} \in \mathbb{R}^{m \times m}$ (or $\mathbb{C}^{m \times m}$) and $\mathbf{V} \in \mathbb{R}^{n \times n}$ (or $\mathbb{C}^{n \times n}$), such that*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \text{ (or } \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H), \quad (2.1)$$

where,

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and $\mathbf{\Sigma} = \text{diag}[\sigma_1, \sigma_2, \dots, \sigma_r]$, its diagonal elements are arranged in the order:

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0, \quad t = \text{rank}(\mathbf{A})$$

The quantity $\sigma_1, \sigma_2, \dots, \sigma_r$ together with $\sigma_{r+1} = \sigma_{r+2} = \dots = \sigma_n = 0$ are called the singular values of matrix \mathbf{A} . The column vector \mathbf{u}_i of matrix \mathbf{U} is called the left singular vector of \mathbf{A} , and the matrix \mathbf{U} is called the left singular matrix. The column vector \mathbf{v}_i of matrix \mathbf{V} is called the right singular vector of \mathbf{A} , and the matrix \mathbf{V} is called the right singular matrix. The proof of Theorem 2.1 can see [4, 5].

The SVD of matrix \mathbf{A} can also be written as:

$$\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^H. \quad (2.2)$$

It can be easily seen that

$$\mathbf{A}\mathbf{A}^H = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^H \quad (2.3)$$

which shows that the singular value σ_i of the $m \times n$ matrix \mathbf{A} is the positive square root of the eigenvalue (these eigenvalues are nonpositive) of the matrix product $\mathbf{A}\mathbf{A}^H$.

The following theorem strictly narrates the singular property of a matrix \mathbf{A} .

Theorem 2.2 Define the singular values of matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ ($m > n$) as $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$.

Then

$$\sigma_k = \min_{\mathbf{E} \in \mathbb{C}^{m \times n}} \left\{ \|\mathbf{E}\|_{\text{spec}} : \text{rank}(\mathbf{A} + \mathbf{E}) \leq (k-1) \right\}, \quad k = 1, 2, \dots, n \quad (2.4)$$

and there is an error matrix which meets $\|\mathbf{E}_k\|_{\text{spec}} = \sigma_k$, so that

$$\text{rank}(\mathbf{A} + \mathbf{E}_k) = r - 1, \quad k = 1, 2, \dots, n.$$

Theorem 2.2 shows that the singular value of a matrix is equal to the spectral norm of the error matrix \mathbf{E}_k which makes the rank of the original matrix reduce one. If the original $n \times n$ matrix \mathbf{A} is square and it has a zero singular value, the spectral norm of error matrix whose rank reduces to one is equal to zero. That is to say, when the original $n \times n$ matrix \mathbf{A} has a zero singular value, the rank of the matrix is $\text{rank}(\mathbf{A}) \leq n - 1$ and the original matrix is not full-rank essentially. So, if a matrix has a zero singular value, the matrix must be singular matrix. Generally speaking, if a rectangle matrix has a zero singular value, then it must not be full column rank or full row rank. This case is called rank-deficient matrix, which is a singular phenomenon with regards to the full-rank matrix.

In the following, we discuss the uniqueness of SVD.

- (1) The number r of nonzero singular values and their values $\sigma_1, \sigma_2, \dots, \sigma_r$ is unique relative to matrix \mathbf{A} .
- (2) If $\text{rank}(\mathbf{A}) = r$, the dimension of the sets of vector $\mathbf{x} \in \mathbb{C}^n$ which meets $\mathbf{A}\mathbf{x} = 0$, namely the zero space of matrix \mathbf{A} , is equal to $n - r$. Thus, one can select orthogonal basis $\{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \dots, \mathbf{v}_n\}$ as the zero space of matrix \mathbf{A} in \mathbb{C}^n . From this point, the subspace $\text{Null}(\mathbf{A})$ of \mathbb{C}^n spanned by column vectors of \mathbf{V} is uniquely determined. However, as long as every vector can constitute the orthogonal basis of this subspace, they can be selected arbitrarily.
- (3) The sets of $\mathbf{y} (\in \mathbb{C}^m)$ which can be denoted as $\mathbf{y} = \mathbf{A}\mathbf{x}$ constitute the image space $\text{Im}\mathbf{A}$ of matrix \mathbf{A} , whose dimension is equal to r . The orthogonal supplement space $(\text{Im}\mathbf{A})^\perp$ of $\text{Im}\mathbf{A}$ is $m-r$ dimensional. Thus, one can select $\{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \dots, \mathbf{u}_m\}$ as the orthogonal basis of $(\text{Im}\mathbf{A})^\perp$. The subspace $(\text{Im}\mathbf{A})^\perp$ of \mathbb{C}^m spanned by the column vectors $\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \dots, \mathbf{u}_m$ of \mathbf{U} is uniquely determined.

- (4) If σ_i is single singular value ($\sigma_i \neq \sigma_j, \forall j \neq i$), \mathbf{v}_i and \mathbf{u}_i is uniquely determined except discrepancy of an angle. That is to say, after \mathbf{v}_i and \mathbf{u}_i multiply $e^{i\theta}$ ($j = \sqrt{-1}$) and θ is real number) at the same time, they are still the right and left singular vectors, respectively.

2.2.2 Properties of SVD

Assume $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$, and $r_A = \text{rank}(\mathbf{A})$, $p = \min\{m, n\}$. The singular values of matrix \mathbf{A} can be arranged as follows: $\sigma_{\max} = \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{p-1} \geq \sigma_p = \sigma_{\min} \geq 0$, and denote by $\sigma_i(\mathbf{B})$ the i th largest singular value of matrix \mathbf{B} . A few properties of SVD can summarized as follows [6]:

- (1) The relationship between the singular values of a matrix and the ones of its submatrix.

Theorem 2.3 (interlacing theorem for singular values). *Assume $\mathbf{A} \in \mathbb{R}^{m \times n}$, and its singular values satisfy $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$, where $r = \min\{m, n\}$. If $\mathbf{B} \in \mathbb{R}^{p \times q}$ is a submatrix of \mathbf{A} , and its singular values satisfy $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_{\min\{p, q\}}$, then it holds that*

$$\sigma_i \geq \gamma_i, \quad i = 1, 2, \dots, \min\{p, q\} \quad (2.5)$$

and

$$\gamma_i \geq \sigma_{i+(m-p)+(n-q)}, \quad i \leq \min\{p+q-m, p+q-n\}. \quad (2.6)$$

From Theorem 2.3, it holds that: If $\mathbf{B} \in \mathbb{R}^{m \times (n-1)}$ is a submatrix of $\mathbf{A} \in \mathbb{R}^{m \times n}$ by deleting any column of matrix \mathbf{A} , and their singular values are arranged in non-decreasing order, then it holds that

$$\sigma_1(\mathbf{A}) \geq \sigma_1(\mathbf{B}) \geq \sigma_2(\mathbf{A}) \geq \sigma_2(\mathbf{B}) \geq \dots \geq \sigma_h(\mathbf{A}) \geq \sigma_h(\mathbf{B}) \geq 0, \quad (2.7)$$

where $h = \min\{m, n-1\}$.

If $\mathbf{B} \in \mathbb{R}^{(m-1) \times n}$ is a submatrix of $\mathbf{A} \in \mathbb{R}^{m \times n}$ by deleting any row of matrix \mathbf{A} , and their singular values are arranged as non-decreasing order, then it holds that

$$\sigma_1(\mathbf{A}) \geq \sigma_1(\mathbf{B}) \geq \sigma_2(\mathbf{A}) \geq \sigma_2(\mathbf{B}) \geq \dots \sigma_h(\mathbf{A}) \geq \sigma_h(\mathbf{B}) \geq 0. \quad (2.8)$$

- (2) The relationship between the singular values of a matrix and its norms.
The spectral norm of a matrix \mathbf{A} is equal to its largest singular value, namely,

$$\|\mathbf{A}\|_{\text{spec}} = \sigma_1. \quad (2.9)$$

According to the SVD theorem of matrix and the unitary invariability property of Frobenius norm $\|\mathbf{A}\|_F$ of matrix \mathbf{A} , namely $\|\mathbf{U}^H \mathbf{A} \mathbf{V}\|_F = \|\mathbf{A}\|_F$, it holds that

$$\|\mathbf{A}\|_F = \left[\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right]^{1/2} = \|\mathbf{U}^H \mathbf{A} \mathbf{V}\|_F = \|\mathbf{\Sigma}\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2}. \quad (2.10)$$

That is to say, the Frobenius norm of any matrix is equal to the square root of the sum of the squares of all nonzero singular values of this matrix.

Consider the rank- k approximation of matrix \mathbf{A} and denote it as \mathbf{A}_k , in which $k < r = \text{rank}(\mathbf{A})$. The matrix \mathbf{A}_k is defined as follows:

$$\mathbf{A}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^H, k < r,$$

Then the spectral norm of the difference between \mathbf{A} and any rank(k) matrix \mathbf{B} , and the Frobenius norm of the difference can be written, respectively, as follows:

$$\min_{\text{rank}(\mathbf{B})=r} \|\mathbf{A} - \mathbf{B}\|_{\text{spec}} = \|\mathbf{A} - \mathbf{A}_k\|_{\text{spec}} = \sigma_{k+1}, \quad (2.11)$$

$$\min_{\text{rank}(\mathbf{B})=r} \|\mathbf{A} - \mathbf{B}\|_F^2 = \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sigma_{k+1}^2 + \sigma_{k+2}^2 + \cdots + \sigma_r^2. \quad (2.12)$$

The above properties are the basis of many concepts and applications. For example, the total least squares, data compression, image enhancement, the solution of linear equations, etc., all need to approximate \mathbf{A} using a lower rank matrix.

- (3) The relationship between the singular values of a matrix and its determinant. Define \mathbf{A} as an $n \times n$ square matrix. Since the absolute value of the determinant of a unitary matrix is equal to one, from SVD theorem it holds that

$$|\det(\mathbf{A})| = |\det \mathbf{\Sigma}| = \sigma_1 \sigma_2 \cdots \sigma_n. \quad (2.13)$$

If all σ_i are non-zero, then $|\det(\mathbf{A})| \neq 0$, which means that \mathbf{A} is nonsingular. If at least one $\sigma_i (i > r)$ is equal to zero, then $|\det(\mathbf{A})| = 0$, namely \mathbf{A} is singular.

- (4) The relationship between the singular values of a matrix and its condition number.

For an $m \times n$ matrix \mathbf{A} , its condition number can be defined using SVD as

$$\text{cond}(\mathbf{A}) = \sigma_1/\sigma_p, \quad p = \min\{m, n\}. \quad (2.14)$$

Since $\sigma_1 \geq \sigma_p$, the condition number is a positive number which is equal to or larger than one. Obviously, since there is at least one singular value which meets $\sigma_p = 0$, the condition number of a singular matrix is infinite. When the condition number, though not infinite, is very large, the matrix \mathbf{A} is called to be close to singular. Since the condition number of unitary or orthogonal matrix is equal to one, the unitary or orthogonal matrix is of “ideal condition”. Equation (2.14) can be used to evaluate the condition number.

- (5) Maximal singular value and minimal singular value.

If $m \geq n$, for any matrix $\mathbf{A}_{m \times n}$, it holds that

$$\begin{aligned} \sigma_{\min}(\mathbf{A}) &= \min \left\{ \left(\frac{\mathbf{x}^H \mathbf{A}^H \mathbf{A} \mathbf{x}}{\mathbf{x}^H \mathbf{x}} \right)^{1/2} : \mathbf{x} \neq 0 \right\} \\ &= \min \left\{ (\mathbf{x}^H \mathbf{A}^H \mathbf{A} \mathbf{x})^{1/2} : \mathbf{x}^H \mathbf{x} = 1, \mathbf{x} \in \mathbb{C}^n \right\} \end{aligned} \quad (2.15)$$

and

$$\begin{aligned} \sigma_{\max}(\mathbf{A}) &= \max \left\{ \left(\frac{\mathbf{x}^H \mathbf{A}^H \mathbf{A} \mathbf{x}}{\mathbf{x}^H \mathbf{x}} \right)^{1/2} : \mathbf{x} \neq 0 \right\} \\ &= \max \left\{ (\mathbf{x}^H \mathbf{A}^H \mathbf{A} \mathbf{x})^{1/2} : \mathbf{x}^H \mathbf{x} = 1, \mathbf{x} \in \mathbb{C}^n \right\}. \end{aligned} \quad (2.16)$$

- (6) The relationship between the singular values and eigenvalues.

Suppose that the eigenvalues of an $n \times n$ symmetrical square matrix \mathbf{A} are $\lambda_1, \lambda_2, \dots, \lambda_n$ ($|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$), and its singular values are $\sigma_1, \sigma_2, \dots, \sigma_n$ ($\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$). Then $\sigma_i \geq |\lambda_i| \geq \sigma_n$ ($i = 1, 2, \dots, n$) and $\text{cond}(\mathbf{A}) \geq |\lambda_1|/|\lambda_n|$.

2.3 Eigenvalue Decomposition

2.3.1 Eigenvalue Problem and Eigen Equation

The basic problem of the eigenvalue can be stated as follows. Given an $n \times n$ matrix \mathbf{A} , determine a scalar λ such that the following algebra equation

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \quad \mathbf{u} \neq 0 \quad (2.17)$$

has an $n \times 1$ nonzero solution. The scalar λ is called as an eigenvalue of matrix \mathbf{A} , and the vector \mathbf{u} is called as the eigenvector associated with λ . Since the eigenvalue

λ and eigenvector \mathbf{u} appear in couples, (λ, \mathbf{u}) is usually called as an eigen pair of matrix \mathbf{A} . Although the eigenvalues can be zeros, the eigenvectors cannot be zero.

In order to determine a nonzero vector \mathbf{u} , Eq. (2.17) can be modified as

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{u} = \mathbf{0}. \quad (2.18)$$

The above equation should come into existence for any vector \mathbf{u} , so the unique condition under which Eq. (2.18) has a nonzero solution $\mathbf{u} = \mathbf{0}$ is that the determinant of matrix $\mathbf{A} - \lambda \mathbf{I}$ is equal to zero, namely

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0. \quad (2.19)$$

Thus, the solution of the eigenvalue problem consists of the following two steps:

- (1) Solve all scalar λ (eigenvalues) which make the matrix $\mathbf{A} - \lambda \mathbf{I}$ singular.
- (2) Given an eigenvalue λ which makes $\mathbf{A} - \lambda \mathbf{I}$ singular, and to solve all nonzero vectors which meets $(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$, i.e., the eigenvectors corresponding to λ .

According to the relationship between the singular values of a matrix and its determinant, a matrix is singular if and only if $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$, namely

$$(\mathbf{A} - \lambda \mathbf{I}) \text{ singular} \Leftrightarrow \det(\mathbf{A} - \lambda \mathbf{I}) = 0. \quad (2.20)$$

The matrix $(\mathbf{A} - \lambda \mathbf{I})$ is called as the eigen matrix of \mathbf{A} . When \mathbf{A} is an $n \times n$ matrix, spreading the left side determinant of Eq. (2.20) can obtain a polynomial equation (power- n), namely

$$\alpha_0 + \alpha_1 \lambda + \cdots + \alpha_{n-1} \lambda^{n-1} + (-1)^n \lambda^n = 0, \quad (2.21)$$

which is called as the eigen equation of matrix \mathbf{A} . The polynomial $\det(\mathbf{A} - \lambda \mathbf{I})$ is called as the eigen polynomial.

2.3.2 Eigenvalue and Eigenvector

In the following, we list some major properties about the eigenvalues and eigenvector of a matrix \mathbf{A} .

Several important terms about the eigenvalues and eigenvectors [6]:

- (1) The eigenvalue λ of a matrix \mathbf{A} is called as having algebraic multiplicity μ , if λ is a μ -repeated root of the eigen equation $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$.
- (2) If the algebraic multiplicity of eigenvalue λ is equal to one, the eigenvalue is called as single eigenvalue. Non-single eigenvalues are called as multiple eigenvalues.
- (3) The eigenvalue λ of a matrix \mathbf{A} is called as having geometric multiplicity γ , if the number of linear independent eigenvectors associated with λ is equal to γ .

- (4) An eigenvalue is called half-single eigenvalue if its algebraic multiplicity is equal to geometric multiplicity. Not half-single eigenvalues are called as wane eigenvalues.
- (5) If matrix $\mathbf{A}_{n \times n}$ is a general complex matrix and λ is its eigenvalue, the vector \mathbf{v} which meets $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ is called as the right eigenvector associated with the eigenvalue λ , and the eigenvector \mathbf{u} which meets $\mathbf{u}^H\mathbf{A} = \lambda\mathbf{u}^H$ is called as the left eigenvector associated with the eigenvalue λ . If \mathbf{A} is Hermitian matrix and all its eigenvalues are real number, then it holds that $\mathbf{v} = \mathbf{u}$, that is to say, the left and right eigenvectors of a Hermitian matrix are the same.

Some important properties can be summarized as follows:

- (1) Matrix $\mathbf{A} (\in \mathbb{R}^{n \times n})$ has n eigenvalues, of which the multiple eigenvalues are computed according to their multiplicity.
- (2) If \mathbf{A} is a real symmetrical matrix or Hermitian matrix, all its eigenvalues are real numbers.
- (3) If $\mathbf{A} = \text{diag}(a_{11}, a_{22}, \dots, a_{nn})$, its eigenvalues are $a_{11}, a_{22}, \dots, a_{nn}$; If \mathbf{A} is a trigonal matrix, its diagonal elements are all its eigenvalues.
- (4) For $\mathbf{A} (\in \mathbb{R}^{n \times n})$, if λ is the eigenvalue of matrix \mathbf{A} , λ is also the eigenvalue of matrix \mathbf{A}^T . If λ is the eigenvalue of matrix \mathbf{A} , λ^* is the eigenvalue of matrix \mathbf{A}^H . If λ is the eigenvalue of matrix \mathbf{A} , $\lambda + \sigma^2$ is the eigenvalue of matrix $\mathbf{A} + \sigma^2\mathbf{I}$. If λ is the eigenvalue of matrix \mathbf{A} , $1/\lambda$ is the eigenvalue of matrix \mathbf{A}^{-1} .
- (5) All eigenvalues of matrix $\mathbf{A}^2 = \mathbf{A}$ are either 0 or 1.
- (6) If \mathbf{A} is a real orthogonal matrix, all its eigenvalues are on the unit circle.
- (7) If a matrix is singular, at least one of its eigenvalues is equal to zero.
- (8) The sum of all the eigenvalues is equal to its trace, namely $\sum_{i=1}^n \lambda_i = \text{tr}(\mathbf{A})$.
- (9) The nonzero eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ associated with different eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ are linearly independent.
- (10) If matrix $\mathbf{A} (\in \mathbb{R}^{n \times n})$ has r nonzero eigenvalues, then it holds that $\text{rank}(\mathbf{A}) \geq r$; If zero is a non-multiple eigenvalue, then $\text{rank}(\mathbf{A}) \geq n - 1$; If $\text{rank}(\mathbf{A} - \lambda\mathbf{I}) \geq n - 1$, then λ is an eigenvalue of matrix \mathbf{A} .
- (11) The product of all eigenvalues of matrix \mathbf{A} is equal to the determinant of matrix \mathbf{A} , namely $\prod_{i=1}^n \lambda_i = \det(\mathbf{A}) = |\mathbf{A}|$.
- (12) A Hermitian matrix \mathbf{A} is positive definite (or positive semi-definite), if and only if all its eigenvalues are positive (or non-negative).
- (13) If the eigenvalues of matrix \mathbf{A} are different, then one can find a similar matrix such that $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{D}$ (diagonal matrix) and the diagonal elements of \mathbf{D} are the eigenvalues of matrix \mathbf{A} .
- (14) (Cayley–Hamilton Theorem) : If $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of an $n \times n$ matrix \mathbf{A} , then $\prod_{i=1}^n (\mathbf{A} - \lambda_i\mathbf{I}) = \mathbf{0}$.

- (15) It is not possible that the geometric multiplicity of any eigenvalue λ of an $n \times n$ matrix \mathbf{A} is larger than its algebraic multiplicity.
- (16) If λ is an eigenvalue of an $n \times n$ matrix \mathbf{A} and an $n \times n$ matrix \mathbf{B} is not singular, then λ is also an eigenvalue of $\mathbf{B}^{-1}\mathbf{A}\mathbf{B}$. However, the corresponding eigenvectors are usually different. If λ is an eigenvalue of an $n \times n$ matrix \mathbf{A} and an $n \times n$ matrix \mathbf{B} is a unitary matrix, then λ is also an eigenvalue of $\mathbf{B}^H\mathbf{A}\mathbf{B}$. However, the corresponding eigenvectors are usually different. If λ is an eigenvalue of an $n \times n$ matrix \mathbf{A} and an $n \times n$ matrix \mathbf{B} is a orthogonal matrix, then λ is also an eigenvalue of $\mathbf{B}^T\mathbf{A}\mathbf{B}$. However, the corresponding eigenvectors are usually different.
- (17) The largest eigenvalue of an $n \times n$ matrix $\mathbf{A} = [a_{ij}]$ is less than or equal to the maximal of the sum of all the column elements of this matrix, namely $\lambda_{\max} \leq \max_i \sum_{j=1}^n a_{ij}$.
- (18) The eigenvalues of autocorrelation matrix $\mathbf{R} = E\{\mathbf{x}(t)\mathbf{x}^H(t)\}$ of stochastic vector $\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_n(t)]^T$ is within the maximal power of signal $P_{\max} = \max_i E\{|x_i(t)|^2\}$ and its minimal power $P_{\min} = \min_i E\{|x_i(t)|^2\}$, namely $P_{\min} \leq \lambda_i \leq P_{\max}$.
- (19) The spread of eigenvalues in autocorrelation matrix \mathbf{R} of a stochastic vector $\mathbf{x}(t)$ is $\kappa(\mathbf{R}) = \lambda_{\max}/\lambda_{\min}$.
- (20) If $|\lambda_i| < 1, i = 1, 2, \dots, n$, the matrix $\mathbf{A} \pm \mathbf{I}_n$ is nonsingular. $|\lambda_i| < 1, i = 1, 2, \dots, n$, is equivalent to the case in which the roots of $\det(\mathbf{A} - z\mathbf{I}_n) = 0$ is not on or at the interior of the unit circle.
- (21) For $m \times n (n \geq m)$ matrix \mathbf{A} and $n \times m$ matrix \mathbf{B} , if λ is an eigenvalue of the product \mathbf{AB} , then λ is also an eigenvalue of the product \mathbf{BA} . If $\lambda \neq 0$ is an eigenvalue of the product \mathbf{BA} , then λ is also an eigenvalue of the product \mathbf{AB} . If $\lambda_1, \lambda_2, \dots, \lambda_m$ are eigenvalues of the product \mathbf{AB} , then the eigenvalues of matrix product \mathbf{BA} are $\lambda_1, \lambda_2, \dots, \lambda_m, 0, \dots, 0$.
- (22) If the eigenvalue of matrix \mathbf{A} is λ , then the eigenvalue of matrix polynomial $f(\mathbf{A}) = \mathbf{A}^n + c_1\mathbf{A}^{n-1} + \dots + c_{n-1}\mathbf{A} + c_n\mathbf{I}$ is $f(\lambda) = \lambda^n + c_1\lambda^{n-1} + \dots + c_{n-1}\lambda + c_n$.
- (23) If λ is an eigenvalue of matrix \mathbf{A} , then the eigenvalue of matrix exponential function $e^{\mathbf{A}}$ is e^λ .

Properties of an eigen pair which consists of an eigenvalue λ and its associated eigenvector \mathbf{u} can be summarized as follows:

- (1) If (λ, \mathbf{u}) is an eigen pair of matrix \mathbf{A} , then $(c\lambda, \mathbf{u})$ is an eigen pair of matrix $c\mathbf{A}$, where c is a nonzero constant.
- (2) If (λ, \mathbf{u}) is an eigen pair of matrix \mathbf{A} , then $(\lambda, c\mathbf{u})$ is an eigen pair of matrix \mathbf{A} , where c is a nonzero constant.
- (3) If $(\lambda_i, \mathbf{u}_i)$ and $(\lambda_j, \mathbf{u}_j)$ are eigen pairs of matrix \mathbf{A} and $\lambda_i \neq \lambda_j$, then the eigenvector \mathbf{u}_i and \mathbf{u}_j are linearly independent.

- (4) The eigenvectors of an Hermitian matrix associated with different eigenvalues are mutual orthogonal to each other, namely $\lambda_i \neq \lambda_j \Rightarrow \mathbf{u}_i^H \mathbf{u}_j = 0$.
- (5) If λ is an eigenvalue of matrix \mathbf{A} and the vectors \mathbf{u}_1 and \mathbf{u}_2 are the eigenvectors associated with λ , then $c_1 \mathbf{u}_1 + c_2 \mathbf{u}_2$ is also an eigenvector of matrix \mathbf{A} associated with the eigenvalue λ , in which c_1 and c_2 are constants and at least one of them is not zero.
- (6) If (λ, \mathbf{u}) is an eigen pair of matrix \mathbf{A} and $\alpha_1, \alpha_2, \dots, \alpha_p$ are complex constants, then $f(\lambda) = \alpha_0 + \alpha_1 \lambda + \dots + \alpha_p \lambda^p$ is the eigenvalue of matrix polynomial $f(\mathbf{A}) = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A} + \dots + \alpha_p \mathbf{A}^p$, and the associated eigenvector is still \mathbf{u} .
- (7) If (λ, \mathbf{u}) is an eigen pair of matrix \mathbf{A} , then (λ^k, \mathbf{u}) is an eigen pair of matrix \mathbf{A}^k .
- (8) If (λ, \mathbf{u}) is an eigen pair of matrix \mathbf{A} , then (e^λ, \mathbf{u}) is an eigen pair of matrix exponential function $e^{\mathbf{A}}$.
- (9) If $\lambda(\mathbf{A})$ and $\lambda(\mathbf{B})$ are eigenvalues of matrices \mathbf{A} and \mathbf{B} , respectively, and $\mathbf{u}(\mathbf{A})$ and $\mathbf{u}(\mathbf{B})$ are their associated eigenvectors, then $\lambda(\mathbf{A})\lambda(\mathbf{B})$ is an eigenvalue of matrix Kronecker product $\mathbf{A} \otimes \mathbf{B}$ with $\mathbf{u}(\mathbf{A}) \otimes \mathbf{u}(\mathbf{B})$ being the associated eigenvector, and $\lambda(\mathbf{A})$ and $\lambda(\mathbf{B})$ are the eigenvalues of matrix direct sum $\mathbf{A} \oplus \mathbf{B}$ with $\begin{bmatrix} \mathbf{u}(\mathbf{A}) \\ \mathbf{0} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{0} \\ \mathbf{u}(\mathbf{B}) \end{bmatrix}$ being the associated eigenvectors, respectively.
- (10) If an $n \times n$ matrix \mathbf{A} has n linearly independent eigenvectors, then its ED is $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{-1}$, where the $n \times n$ matrix \mathbf{U} consists of n eigenvectors of matrix \mathbf{A} , and the diagonal elements of the $n \times n$ diagonal matrix $\mathbf{\Sigma}$ are the eigenvalues of matrix \mathbf{A} .

The SVD problem of a matrix \mathbf{A} can be transformed into its ED problem to solve, and there are two methods to realize this.

Method 2.1 The nonzero singular values of matrix $\mathbf{A}_{m \times n}$ are the positive square root of nonzero eigenvalue λ_i of $m \times m$ matrix $\mathbf{A}\mathbf{A}^T$ or $n \times n$ matrix $\mathbf{A}^T\mathbf{A}$, and the left singular vector \mathbf{u}_i and right singular vector \mathbf{v}_i of matrix \mathbf{A} associated with σ_i are the eigenvectors of matrix $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ associated with nonzero eigenvalue λ_i , respectively.

Method 2.2 The SVD of matrix $\mathbf{A}_{m \times n}$ can be transformed into the ED of $(m+n) \times (m+n)$ augmented matrix $\begin{bmatrix} \mathbf{O} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{O} \end{bmatrix}$.

The following theorem holds for the eigenvalues of matrix sum $\mathbf{A} + \mathbf{B}$.

Theorem 2.4 (Wely theorem): Suppose that $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m \times n}$ are Hermitian matrices, and their eigenvalues are arranged as an increasing order, namely,

$$\lambda_1(\mathbf{A}) \leq \lambda_2(\mathbf{A}) \leq \dots \leq \lambda_n(\mathbf{A}),$$

$$\lambda_1(\mathbf{B}) \leq \lambda_2(\mathbf{B}) \leq \dots \leq \lambda_n(\mathbf{B}),$$

$$\lambda_1(\mathbf{A} + \mathbf{B}) \leq \lambda_2(\mathbf{A} + \mathbf{B}) \leq \dots \leq \lambda_n(\mathbf{A} + \mathbf{B}),$$

Then,

$$\lambda_i(\mathbf{A} + \mathbf{B}) \geq \begin{cases} \lambda_i(\mathbf{A}) + \lambda_1(\mathbf{B}) \\ \lambda_{i-1}(\mathbf{A}) + \lambda_2(\mathbf{B}) \\ \vdots \\ \lambda_1(\mathbf{A}) + \lambda_i(\mathbf{B}) \end{cases} \quad (2.22)$$

and

$$\lambda_i(\mathbf{A} + \mathbf{B}) \leq \begin{cases} \lambda_i(\mathbf{A}) + \lambda_n(\mathbf{B}) \\ \lambda_{i+1}(\mathbf{A}) + \lambda_{n-1}(\mathbf{B}) \\ \vdots \\ \lambda_n(\mathbf{A}) + \lambda_i(\mathbf{B}). \end{cases} \quad (2.23)$$

where $i = 1, 2, \dots, n$.

Especially, when \mathbf{A} is a real symmetric matrix, and $\mathbf{B} = \mathbf{a}\mathbf{z}\mathbf{z}^T$, the interlace theorem in the following holds.

Theorem 2.5 (Interlacing eigenvalue theorem): *Suppose that $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a symmetric matrix, and its eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, meet $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, and let $\mathbf{z} \in \mathbb{R}^n$ be a vector satisfying $\|\mathbf{z}\| = 1$. Suppose that a is a real number and the eigenvalues of matrix $\mathbf{A} + \mathbf{a}\mathbf{z}\mathbf{z}^T$ meet $\zeta_1 \geq \zeta_2 \geq \dots \geq \zeta_n$, then it holds that*

$$\zeta_1 \geq \lambda_1 \geq \zeta_2 \geq \lambda_2 \geq \dots \geq \zeta_n \geq \lambda_n, \quad a > 0 \quad (2.24)$$

or

$$\lambda_1 \geq \zeta_1 \geq \lambda_2 \geq \zeta_2 \geq \dots \geq \lambda_n \geq \zeta_n, \quad a < 0 \quad (2.25)$$

and whether $a > 0$ or $a < 0$, it holds that

$$\sum_{i=1}^n (\zeta_i - \lambda_i) = a. \quad (2.26)$$

2.3.3 Eigenvalue Decomposition of Hermitian Matrix

All the discussions on eigenvalues and eigenvectors in the above hold for general matrices, and they do not require the matrices to be real symmetric or complex conjugate symmetric. However, in the statistical and information science, one usually encounter real symmetric or Hermitian (complex conjugate symmetric) matrices. For example, the autocorrelation matrix of a real measurement data vector $\mathbf{R} = E\{\mathbf{x}(t)\mathbf{x}^T(t)\}$ is real symmetric, while the autocorrelation matrix of a complex measurement data vector $\mathbf{R} = E\{\mathbf{x}(t)\mathbf{x}^H(t)\}$ is Hermitian. On the other hand, since a real symmetric matrix is a special case of Hermitian matrix and the eigenvalues and eigenvectors of a Hermitian matrix have a series of important properties, and it is necessary to discuss individually the eigen analysis of Hermitian matrix.

1. Eigenvalue and Eigenvector of Hermitian matrix.

Some important properties of eigenvalues and eigenvectors of Hermitian matrices can be summarized as follows:

- (1) The eigenvalues of an Hermitian matrix \mathbf{A} must be a real number.
- (2) Let (λ, \mathbf{u}) be an eigen pair of an Hermitian matrix \mathbf{A} . If \mathbf{A} is invertible, then $(1/\lambda, \mathbf{u})$ is an eigen pair of matrix \mathbf{A}^{-1} .
- (3) If λ_k is a multiple eigenvalue of Hermitian matrix $\mathbf{A}^H = \mathbf{A}$, and its multiplicity is m_k , then $\text{rank}(\mathbf{A} - \lambda_k \mathbf{I}) = n - m_k$.
- (4) Any Hermitian matrix \mathbf{A} is diagonalizable, namely $\mathbf{U}^{-1}\mathbf{A}\mathbf{U} = \mathbf{\Sigma}$.
- (5) All the eigenvectors of an Hermitian matrix are linearly independent, and they are mutual orthogonal, namely the eigen matrix $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n]$ is a unitary matrix and it meets $\mathbf{U}^{-1} = \mathbf{U}^H$.
- (6) From property (5), it holds that $\mathbf{U}^H\mathbf{A}\mathbf{U} = \mathbf{\Sigma} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ or $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$, which can be rewritten as: $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^H$. This is called the spectral decomposition of a Hermitian matrix.
- (7) The spread formula of the inverse of an Hermitian matrix \mathbf{A} is

$$\mathbf{A}^{-1} = \sum_{i=1}^n \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^H \quad (2.27)$$

Thus, if one know the eigen decomposition of an Hermitian matrix \mathbf{A} , then one can directly obtain the inverse matrix \mathbf{A}^{-1} using the above formula.

- (8) For two $n \times n$ Hermitian matrices \mathbf{A} and \mathbf{B} , there exists a unitary matrix so that $\mathbf{P}^H\mathbf{A}\mathbf{P}$ and $\mathbf{P}^H\mathbf{B}\mathbf{P}$ are both diagonal if and only if $\mathbf{AB} = \mathbf{BA}$.
- (9) For two $n \times n$ non-negative definite Hermitian matrices \mathbf{A} and \mathbf{B} , there exists a nonsingular matrix \mathbf{P} so that $\mathbf{P}^H\mathbf{A}\mathbf{P}$ and $\mathbf{P}^H\mathbf{B}\mathbf{P}$ are both diagonal.

2. Some properties of Hermitian matrix.

The ED of an Hermitian matrix \mathbf{A} can be written as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$, where \mathbf{U} is a unitary matrix and it meets $\mathbf{U}^H\mathbf{U} = \mathbf{U}\mathbf{U}^H = \mathbf{I}$.

From the property of determinant and trace of a matrix, it holds that

$$\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{U}\mathbf{\Sigma}\mathbf{U}^H) = \text{tr}(\mathbf{U}^H\mathbf{U}\mathbf{\Sigma}) = \text{tr}(\mathbf{\Sigma}) = \sum_{i=1}^n \lambda_i, \quad (2.28)$$

$$\det(\mathbf{A}) = \det(\mathbf{U}) \det(\mathbf{\Sigma}) \det(\mathbf{U}^H) = \prod_{i=1}^n \lambda_i. \quad (2.29)$$

For a positive definite Hermitian matrix \mathbf{A} , its inverse \mathbf{A}^{-1} exists and can be written as

$$\mathbf{A}^{-1} = \mathbf{U} \text{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_u^{-1}) \mathbf{U}^H. \quad (2.30)$$

Let z_A be the number of zero eigenvalues of matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$, then

$$\text{rank}(\mathbf{A}) = n - z_n, \quad (2.31)$$

That is to say, the rank of a Hermitian matrix is equal to the number of its nonzero eigenvalues.

3. Solving for maximal or minimal eigenvalue of Hermitian matrix.

In signal processing, one usually needs to compute the maximal or minimal eigenvalue of a Hermitian matrix \mathbf{A} . The power iteration method is a method for such purposes.

Select some initial vector $\mathbf{x}(0)$, and iteratively repeat the following linear equation

$$\mathbf{y}(k+1) = \mathbf{A}\mathbf{x}(k) \quad (2.32)$$

to obtain $\mathbf{y}(k+1)$, then normalize it. It holds that

$$\mathbf{x}(k+1) = \frac{\mathbf{y}(k+1)}{\sigma_{k+1}}, \quad (2.33)$$

$$\sigma_{k+1} = \mathbf{y}^H(k+1)\mathbf{y}(k+1). \quad (2.34)$$

The iterative procedure continues until the vector \mathbf{x}_k converges. The σ_k obtained at the last iteration is the maximal eigenvalue, and the \mathbf{x}_k is its associated eigenvector. Only if the initial vector $\mathbf{x}(0)$ is not orthogonal to the eigenvector associated with the maximal eigenvalue, the convergence can be guaranteed.

If one needs to compute the minimal eigenvalue and its associated eigenvector, use $\mathbf{y}(k+1) = \mathbf{A}^{-1}\mathbf{x}(k)$, i.e., the iterative linear equation is $\mathbf{A}\mathbf{y}(k+1) = \mathbf{x}(k)$.

By combining the power iteration method and shrink mapping method, one can compute all eigenvalues and the associated eigenvectors of a Hermitian matrix \mathbf{A} . Suppose that one has obtained some eigenvalue σ using the power iteration method. The first step corresponds to the first maximal eigenvalue and uses the shrink mapping method to eliminate the eigenvalue. Then matrix \mathbf{A}_k ($\text{rank} \mathbf{A}_k = k$) is changed into matrix \mathbf{A}_{k-1} ($\text{rank} \mathbf{A}_{k-1} = k - 1$). Thus, the maximal eigenvalue of matrix \mathbf{A}_{k-1} is the residual maximal eigenvalue of matrix \mathbf{A}_k , which is smaller than σ . It should be noted that the k th step corresponds to the k th maximal eigenvalue. New matrix can be obtained by using the above idea and the following spectral decomposition formula:

$$(\mathbf{A}_k - \sigma \mathbf{x} \mathbf{x}^H) = \mathbf{A}_{k-1}.$$

Repeat the above procedure, one can compute all eigenvalues of matrix \mathbf{A} in turn.

2.3.4 Generalized Eigenvalue Decomposition

Let \mathbf{A} and \mathbf{B} both be $n \times n$ square matrices, and they constitute a matrix pencil or matrix pair, written as (\mathbf{A}, \mathbf{B}) . Now we consider the following generalized eigenvalue problem. That is, to compute all scalar λ such that

$$\mathbf{A}\mathbf{u} = \lambda \mathbf{B}\mathbf{u} \quad (2.35)$$

has nonzero solution $\mathbf{u} \neq 0$, where the scalar λ and the nonzero vector \mathbf{u} are called the generalized eigenvalue and the generalized eigenvector of matrix pencil (\mathbf{A}, \mathbf{B}) , respectively. A generalized eigenvalue and its associated generalized eigenvector are called generalized eigen pair, written as (λ, \mathbf{u}) . Equation (2.35) is also called the generalized eigen equation. It is obvious that the eigenvalue problem is a special case when the matrix pencil is chosen as (\mathbf{A}, \mathbf{I}) .

Theorem 2.6 $\lambda \in \mathbb{C}$ and $\mathbf{u} \in \mathbb{C}^n$ are respectively the generalized eigenvalue and the associated generalized eigenvector of matrix pencil $(\mathbf{A}, \mathbf{B})_{n \times n}$ if and only if:

- (1) $\det(\mathbf{A} - \lambda \mathbf{B}) = 0$.
- (2) $\mathbf{u} \in \text{Null}(\mathbf{A} - \lambda \mathbf{B})$, and $\mathbf{u} \neq 0$.

In the natural science, sometimes it is necessary to discuss the eigenvalue problem of the generalized matrix pencil.

Suppose that $n \times n$ square matrices \mathbf{A} and \mathbf{B} are both Hermitian, and \mathbf{B} is positive definite. Then (\mathbf{A}, \mathbf{B}) is called the regularized matrix pencil.

The eigenvalue problem of regularized matrix pencil is similar to the one of Hermitian matrix.

Theorem 2.7 If $\lambda_1, \lambda_2, \dots, \lambda_n$ are the generalized eigenvalues of a regularized matrix pencil (\mathbf{A}, \mathbf{B}) , then

- (1) there exists a matrix $\mathbf{X} \in \mathbb{C}^{n \times n}$, so that

$$\mathbf{X}\mathbf{B}\mathbf{X}^H = \mathbf{I}_n, \quad \mathbf{X}\mathbf{A}\mathbf{X}^H = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n),$$

or equivalently

$$\mathbf{X}^H\mathbf{B}\mathbf{X} = \mathbf{I}_n, \quad \mathbf{A}\mathbf{X} = \mathbf{B}\mathbf{X}\mathbf{A},$$

where $\mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$.

- (2) all generalized eigenvalues are real numbers, i.e., $\lambda_i \in \mathbb{R}, i = 1, 2, \dots, n$.
 (3) Denote $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$. Then it holds that

$$\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{B}\mathbf{x}_i, \quad i = 1, 2, \dots, n.$$

$$\mathbf{x}_i^H\mathbf{B}\mathbf{x}_j = \delta_{ij}, \quad i, j = 1, 2, \dots, n.$$

where δ_{ij} is the Kronecker δ function.

Some properties of the generalized eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ can be summarized as follows, see [7, pp. 176–177]:

- (1) If we interchange matrices \mathbf{A} and \mathbf{B} , then the generalized eigenvalue will be its reciprocal. However, the generalized eigenvector retain unaltered, i.e.,

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x} \quad \Rightarrow \quad \mathbf{B}\mathbf{x} = \frac{1}{\lambda}\mathbf{A}\mathbf{x}.$$

- (2) If matrix \mathbf{B} is nonsingular, then the generalized ED will be simplified to the standard ED

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x} \quad \Rightarrow \quad (\mathbf{B}^{-1}\mathbf{A})\mathbf{x} = \lambda\mathbf{x}.$$

- (3) If matrices \mathbf{A} and \mathbf{B} are both positive definite and Hermitian, then the generalized eigenvalues must be real numbers, and the generalized eigenvectors associated with different generalized values are orthogonal with respect to the positive definite matrices \mathbf{A} and \mathbf{B} , i.e.,

$$\mathbf{x}_i^H\mathbf{A}\mathbf{x}_j = \mathbf{x}_i^H\mathbf{B}\mathbf{x}_j = 0.$$

- (4) If \mathbf{A} and \mathbf{B} are real symmetrical matrices, and \mathbf{B} is positive definite, then the generalized eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ can be changed into the standard eigenvalue problem,

$$(\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-T})(\mathbf{L}^T\mathbf{x}) = \lambda(\mathbf{L}^T\mathbf{x}),$$

where \mathbf{L} is a lower triangular matrix, which is the factor of Cholesky Decomposition $\mathbf{B} = \mathbf{L}\mathbf{L}^T$.

- (5) If \mathbf{A} and \mathbf{B} are real symmetrical and positive definite matrices, then the generalized eigenvalues must be positive.
- (6) If \mathbf{A} is singular, then $\lambda = 0$ must be a generalized eigenvalue.
- (7) If $\tilde{\mathbf{B}} = \mathbf{B} + (1/\alpha)\mathbf{A}$, where α is a nonzero scalar, then the following relationship holds between the generalized eigenvalue $\tilde{\lambda}$ of the modified generalized value problem $\mathbf{A}\mathbf{x} = \tilde{\lambda}\tilde{\mathbf{B}}\mathbf{x}$ and the original generalized eigenvalue λ , i.e.,

$$\frac{1}{\tilde{\lambda}} = \frac{1}{\lambda} + \frac{1}{\alpha}.$$

In the following, we introduce a few generalized ED algorithms for matrix pencil.

We know that if $n \times n$ square matrices \mathbf{A} and \mathbf{B} are both Hermitian, and \mathbf{B} is positive definite, then the generalized ED Eq. (2.35) can be equivalently written as

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \quad (2.36)$$

That is to say, the generalized ED becomes the standard ED of a Hermitian matrix.

The following algorithm uses the shrink mapping to compute the generalized eigen pair (λ, \mathbf{u}) of an $n \times n$ real symmetrical matrix pencil (\mathbf{A}, \mathbf{B}) .

Algorithm 2.1 Lanczos algorithm for generalized ED [8, p. 298].

Step 1 Initialization

Select vector \mathbf{u}_1 whose norm meets $\mathbf{u}_1^H \mathbf{B} \mathbf{u}_1 = 1$, and let $\alpha_1 = 0, \mathbf{z}_0 = \mathbf{u}_0 = 0, \mathbf{z}_1 = \mathbf{B} \mathbf{u}_1$.

Step 2 For $i = 1, 2, \dots, n$, compute

$$\mathbf{u} = \mathbf{A}\mathbf{u}_i - \alpha_i \mathbf{z}_{i-1}$$

$$\beta_i = \langle \mathbf{u}, \mathbf{u}_i \rangle$$

$$\mathbf{u} = \mathbf{u} - \beta_i \mathbf{z}_i$$

$$\mathbf{w} = \mathbf{B}^{-1}\mathbf{u}$$

$$\alpha_{i+1} = \sqrt{\langle \mathbf{w}, \mathbf{u} \rangle}$$

$$\mathbf{u}_{i+1} = \mathbf{w} / \alpha_{i+1}$$

$$\mathbf{z}_{i+1} = \mathbf{u} / \alpha_{i+1}$$

$$\lambda_i = \beta_{i+1} / \alpha_{i+1}.$$

The following is the tangent algorithm for generalized ED of a $n \times n$ symmetric positive definite matrix pencil (\mathbf{A}, \mathbf{B}) , which was proposed by Dramc in 1998 [9].

Algorithm 2.2 Generalized ED of symmetric positive definite matrix pencil.

- Step 1 Compute $\mathbf{A}_A = \text{diag}(A_{11}, A_{22}, \dots, A_{nn})^{-1/2}$, $\mathbf{A}_S = \mathbf{A}_A \mathbf{A} \mathbf{A}_A$ and $\mathbf{B}_1 = \mathbf{A}_A \mathbf{B} \mathbf{A}_A$,
 Step 2 Compute Cholesky Decomposition $\mathbf{R}_A^T \mathbf{R}_A = \mathbf{A}_S$ and $\mathbf{R}_B^T \mathbf{R}_B = \mathbf{B}_1$.
 Step 3 By solving the matrix equation $\mathbf{F} \mathbf{R}_B = \mathbf{A} \mathbf{\Pi}$, compute $\mathbf{F} = \mathbf{A} \mathbf{\Pi} \mathbf{R}_B^{-1}$.
 Step 4 Conduct the SVD $\mathbf{\Sigma} = \mathbf{V} \mathbf{F} \mathbf{U}^T$.
 Step 5 Compute $\mathbf{X} = \mathbf{A}_A \mathbf{\Pi} \mathbf{R}_B^{-1} \mathbf{U}$.

Output: Matrix \mathbf{X} and $\mathbf{\Sigma}$, which meets $\mathbf{A} \mathbf{X} = \mathbf{B} \mathbf{X} \mathbf{\Sigma}^2$.

When matrix \mathbf{B} is singular, the above algorithms will be unstable. The generalized ED algorithm of matrix pencil (\mathbf{A}, \mathbf{B}) under this condition was proposed by Nour-Omid et al. [10], whose main ideas is to make $(\mathbf{A} - \sigma \mathbf{B})$ nonsingular by introducing a shift factor.

Algorithm 2.3 Generalized ED when matrix \mathbf{B} is singular [8, 10], p. 299].

- Step 1 Initialization
 Select the basis vector \mathbf{w} of $\text{Range}[(\mathbf{A} - \sigma \mathbf{B})^{-1} \mathbf{B}]$, compute $\mathbf{z}_1 = \mathbf{B} \mathbf{w}$, $\alpha_1 = \sqrt{\langle \mathbf{w}, \mathbf{z}_1 \rangle}$. Let $\mathbf{u}_0 = 0$.
 Step 2 For $i = 1, 2, \dots, n$, compute

$$\mathbf{u}_i = \mathbf{w} / \alpha_i$$

$$\mathbf{z}_i = (\mathbf{A} - \sigma \mathbf{B})^{-1} \mathbf{w}$$

$$\mathbf{w} = \mathbf{w} - \alpha_i \mathbf{u}_{i-1}$$

$$\beta_i = \langle \mathbf{w}, \mathbf{z}_i \rangle$$

$$\mathbf{z}_{i+1} = \mathbf{B} \mathbf{w}$$

$$\alpha_{i+1} = \sqrt{\langle \mathbf{z}_{i+1}, \mathbf{w} \rangle}$$

2.4 Rayleigh Quotient and Its Characteristics

The quotient of quadratic function of a Hermitian matrix is defined as Rayleigh quotient. As an important quantity in matrix algebra and physics, Rayleigh quotient is a ratio of quadratic functions expressed by eigenvalues and eigenvectors, which has been widely used in many areas such as optimization, signal processing, pattern recognition, and communication.

2.4.1 Rayleigh Quotient

Definition 2.1 The Rayleigh quotient (RQ) of an Hermitian matrix $\mathbf{C} \in \mathbb{C}^{n \times n}$ is a scalar, defined as

$$r(\mathbf{u}) = r(\mathbf{u}, \mathbf{C}) = \frac{\mathbf{u}^H \mathbf{C} \mathbf{u}}{\mathbf{u}^H \mathbf{u}},$$

where \mathbf{u} is a quantity to be selected. The objective is to maximize or minimize the Rayleigh quotient.

The most relevant properties of the RQ are can be summarized as follows:

- ① Homogeneity: $r(\alpha \mathbf{u}, \beta \mathbf{u}) = \beta r(\mathbf{u}, \mathbf{C}) \quad \forall \alpha, \beta \neq 0$.
- ② Translation invariance: $r(\mathbf{u}, \mathbf{C} - \alpha \mathbf{I}) = r(\mathbf{u}, \mathbf{C}) - \alpha$.
- ③ Boundedness: Since \mathbf{u} ranges over all nonzero vectors, $r(\mathbf{u})$ fills a region in the complex plane which is called the field of values of \mathbf{C} . This region is closed, bounded, and convex. If $\mathbf{C} = \mathbf{C}^*$ (selfadjoint matrix), the field of values is the real interval bounded by the extreme eigenvalues.
- ④ Orthogonality: $\mathbf{u} \perp (\mathbf{C} - r(\mathbf{u})\mathbf{I})\mathbf{u}$.
- ⑤ Minimal residual: $\forall \mathbf{u} \neq 0 \wedge \forall \text{ scalar } \mu, \|(\mathbf{C} - r(\mathbf{u})\mathbf{I})\mathbf{u}\| \leq \|(\mathbf{C} - \mu\mathbf{I})\mathbf{u}\|$.

Proposition 2.1 (Stationarity) *Let \mathbf{C} be a real symmetric n -dimensional matrix with eigenvalues $\lambda_n \leq \lambda_{n-1} \leq \dots \leq \lambda_1$ and associated unit eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$. Then it holds that $\lambda_1 = \max r(\mathbf{u}, \mathbf{C})$, $\lambda_n = \min r(\mathbf{u}, \mathbf{C})$. More generally, the critical points and critical values of $r(\mathbf{u}, \mathbf{C})$ are the eigenvectors and eigenvalues of \mathbf{C} .*

Proposition 2.2 (Degeneracy): *The RQ critical points are degenerate because at these points the Hessian matrix is not invertible. Then the RQ is not a Morse function in every open subspace of the domain containing a critical point.*

Furthermore, the following important theorems also holds for RQ.

Courant–Fischer Theorem: Let $\mathbf{C} \in \mathbb{C}^{n \times n}$ be an Hermitian matrix, and its eigenvalues are $\lambda_1 \geq \lambda_2 \geq \dots \leq \lambda_n$, then it holds that for $\lambda_k (1 \leq k \leq n)$:

$$\lambda_k = \min_{S, \dim(S)=n-k+1} \max_{\mathbf{u} \in S, \mathbf{u} \neq 0} \left(\frac{\mathbf{u}^H \mathbf{C} \mathbf{u}}{\mathbf{u}^H \mathbf{u}} \right).$$

The Courant–Fischer Theorem can also written as

$$\lambda_k = \min_{S, \dim(S)=k} \max_{\mathbf{u} \in S, \mathbf{u} \neq 0} \left(\frac{\mathbf{u}^H \mathbf{C} \mathbf{u}}{\mathbf{u}^H \mathbf{u}} \right).$$

2.4.2 Gradient and Conjugate Gradient Algorithm for RQ

If the negative direction of RQ gradient is regarded as the gradient flow of vector \mathbf{x} , e.g.,

$$\dot{\mathbf{x}} = -[\mathbf{C} - r(\mathbf{x})\mathbf{I}]\mathbf{x}$$

then vector \mathbf{x} can be computed iteratively by the following gradient algorithm:

$$\mathbf{x}(k+1) = \mathbf{x}(k) + \mu \dot{\mathbf{x}} = \mathbf{x}(k) - \mu[\mathbf{C} - r(\mathbf{x})\mathbf{I}]\mathbf{x}.$$

It is worth noting that the gradient algorithm of RQ has faster convergence speed than the iterative algorithm of standard RQ.

In the following, the conjugate gradient algorithm for RQ will be introduced, where \mathbf{A} in the RQ is a real symmetric matrix.

Starting from some initial vector, the conjugate gradient algorithm uses the iterative equation, e.g.,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{P}_k \quad (2.37)$$

to update and approach the eigenvector, associated with the minimal or maximal eigenvalue of a symmetric matrix. The real coefficient α_k is

$$\alpha_k = \pm \frac{1}{2D} \left(-B + \sqrt{B^2 - 4CD} \right), \quad (2.38)$$

where “+” is used in the updating of the eigenvector associated with the minimal eigenvalue, and “−” is used in the updating of the eigenvector associated with the maximal eigenvalue. The formulae for parameters D, B, C in the above equations are

$$\left\{ \begin{array}{l} D = P_b(k)P_c(k) - P_a(k)P_d(k) \\ B = P_b(k) - \lambda_k P_d(k) \\ C = P_a(k) - \lambda_k P_c(k) \\ P_a(k) = \mathbf{P}_k^T \mathbf{A} \mathbf{x}_k / (\mathbf{x}_k^T \mathbf{x}_k) \\ P_b(k) = \mathbf{p}_k^T \mathbf{A} \mathbf{p}_k / (\mathbf{x}_k^T \mathbf{x}_k) \\ P_c(k) = \mathbf{p}_k^T \mathbf{x}_k / (\mathbf{x}_k^T \mathbf{x}_k) \\ P_d(k) = \mathbf{p}_k^T \mathbf{p}_k / (\mathbf{x}_k^T \mathbf{x}_k) \\ \lambda_k = r(\mathbf{x}_k) = \mathbf{x}_k^T \mathbf{A} \mathbf{x}_k / (\mathbf{x}_k^T \mathbf{x}_k). \end{array} \right. \quad (2.39)$$

At the $k + 1$ th iteration, the search direction can be selected as

$$\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + b(k)\mathbf{p}_k, \quad (2.40)$$

where $b(-1) = 0$ and \mathbf{r}_{k+1} is the residual vector at the $k + 1$ th iteration. \mathbf{r}_{k+1} and $b(k)$ can be computed, respectively, as

$$\mathbf{r}_{k+1} = -\frac{1}{2} \nabla_x r(\mathbf{x}_{k+1}) = (\lambda_{k+1} \mathbf{x}_{k+1} - \mathbf{A} \mathbf{x}_{k+1}) / (\mathbf{x}_{k+1}^T \mathbf{x}_{k+1}) \quad (2.41)$$

and

$$b(k) = -\frac{\mathbf{r}_{k+1}^T \mathbf{A} \mathbf{p}_k + (\mathbf{r}_{k+1}^T \mathbf{r}_{k+1})(\mathbf{x}_{k+1}^T \mathbf{p}_k)}{\mathbf{p}_k^T (\mathbf{A} \mathbf{p}_k - \lambda_{k+1} \mathbf{I}) \mathbf{p}_k}. \quad (2.42)$$

Equations (2.5)–(2.9) constitute the conjugate gradient algorithm for RQ, which was proposed in [11]. If the updated \mathbf{x}_k is normalized to one and “+” (or “–”) is selected in Eq. (2.6), the above algorithm will obtain the minimal (or maximal) eigenvalue of matrix \mathbf{A} and its associated eigenvectors.

2.4.3 Generalized Rayleigh Quotient

Definition 2.3 Assume that $\mathbf{A} \in \mathbb{C}^{n \times n}$, $\mathbf{B} \in \mathbb{C}^{n \times n}$ are both Hermitian matrices, and \mathbf{B} is positive definite. The generalized RQ or generalized Rayleigh–Ritz of the matrix pencil (\mathbf{A}, \mathbf{B}) is a scalar function, e.g.,

$$r(\mathbf{x}) = \frac{\mathbf{x}^H \mathbf{A} \mathbf{x}}{\mathbf{x}^H \mathbf{B} \mathbf{x}}, \quad (2.43)$$

where \mathbf{x} is a quantity to be selected, and the objective is to maximize or minimize the generalized RQ.

In order to solve for the generalized RQ, define a new vector $\tilde{\mathbf{x}} = \mathbf{B}^{1/2}\mathbf{x}$, where $\mathbf{B}^{1/2}$ is the square root of the positive definite \mathbf{B} . Replace \mathbf{x} by $\mathbf{B}^{-1/2}\tilde{\mathbf{x}}$ in (2.43). Then it holds that

$$r(\tilde{\mathbf{x}}) = \frac{\tilde{\mathbf{x}}^H (\mathbf{B}^{-1/2})^H \mathbf{A} (\mathbf{B}^{-1/2})^H \tilde{\mathbf{x}}}{\tilde{\mathbf{x}}^H \tilde{\mathbf{x}}}, \quad (2.44)$$

which shows that the generalized RQ of matrix pencil (\mathbf{A}, \mathbf{B}) is equivalent to the RQ of matrix product $(\mathbf{B}^{-1/2})^H \mathbf{A} (\mathbf{B}^{-1/2})^H$. From the Rayleigh–Ritz theorem, it is clear that when vector $\tilde{\mathbf{x}}$ is the eigenvector associated with the smallest eigenvalue λ_{\min} of matrix product $(\mathbf{B}^{-1/2})^H \mathbf{A} (\mathbf{B}^{-1/2})^H$, the generalized RQ obtains λ_{\min} . And if vector $\tilde{\mathbf{x}}$ is the eigenvector associated with the largest eigenvalue λ_{\max} of matrix product $(\mathbf{B}^{-1/2})^H \mathbf{A} (\mathbf{B}^{-1/2})^H$, the generalized RQ obtains λ_{\max} .

In the following, we review the eigen decomposition of matrix product $(\mathbf{B}^{-1/2})^H \mathbf{A} (\mathbf{B}^{-1/2})^H$, e.g.,

$$(\mathbf{B}^{-1/2})^H \mathbf{A} (\mathbf{B}^{-1/2})^H \tilde{\mathbf{x}} = \lambda \tilde{\mathbf{x}}. \quad (2.45)$$

If $\mathbf{B} = \sum_{i=1}^n \beta_i \mathbf{v}_i \mathbf{v}_i^H$ is an eigen decomposition of matrix \mathbf{B} , then

$$\mathbf{B}^{1/2} = \sum_{i=1}^n \sqrt{\beta_i} \mathbf{v}_i \mathbf{v}_i^H$$

and $\mathbf{B}^{1/2} \mathbf{B}^{1/2} = \mathbf{B}$. Since matrix $\mathbf{B}^{1/2}$ and $\mathbf{B}^{-1/2}$ have the same eigenvectors and their eigenvalues are reciprocals to each other, then it follows that

$$\mathbf{B}^{-1/2} = \sum_{i=1}^n \frac{1}{\sqrt{\beta_i}} \mathbf{v}_i \mathbf{v}_i^H,$$

which shows that $\mathbf{B}^{-1/2}$ is also an Hermitian matrix, e.g., $(\mathbf{B}^{-1/2})^H = \mathbf{B}^{-1/2}$.

Premultiply both sides of (2.45) by $\mathbf{B}^{-1/2}$, and use $(\mathbf{B}^{-1/2})^H = \mathbf{B}^{-1/2}$, then it holds that

$$\mathbf{B}^{-1} \mathbf{A} \mathbf{B}^{-1/2} \tilde{\mathbf{x}} = \lambda \mathbf{B}^{-1/2} \tilde{\mathbf{x}}$$

or

$$\mathbf{B}^{-1} \mathbf{A} \mathbf{x} = \lambda \mathbf{x}.$$

Since $\mathbf{x} = \mathbf{B}^{-1/2}\tilde{\mathbf{x}}$, thus the eigen decomposition of matrix product $(\mathbf{B}^{-1/2})^H \mathbf{A} (\mathbf{B}^{-1/2})^H$ is equivalent to the one of matrix $\mathbf{B}^{-1}\mathbf{A}$. The eigen decomposition of matrix $\mathbf{B}^{-1}\mathbf{A}$ is the generalized eigenvalue decompositions of matrix pencil (\mathbf{A}, \mathbf{B}) . Thus, the conditions for the maximum and minimum of generalized RQ are

$$r(\mathbf{x}) = \frac{\mathbf{x}^H \mathbf{A} \mathbf{x}}{\mathbf{x}^H \mathbf{B} \mathbf{x}} = \lambda_{\max}, \quad \mathbf{A} \mathbf{x} = \lambda_{\max} \mathbf{B} \mathbf{x},$$

$$r(\mathbf{x}) = \frac{\mathbf{x}^H \mathbf{A} \mathbf{x}}{\mathbf{x}^H \mathbf{B} \mathbf{x}} = \lambda_{\min}, \quad \mathbf{A} \mathbf{x} = \lambda_{\min} \mathbf{B} \mathbf{x}.$$

That is to say, to maximize the generalized RQ, vector \mathbf{x} must be the eigenvector associated with the largest generalized eigenvalue λ_{\max} of matrix pencil (\mathbf{A}, \mathbf{B}) . And to minimize the generalized RQ, vector \mathbf{x} must be the eigenvector associated with the smallest generalized eigenvalue λ_{\min} of matrix pencil (\mathbf{A}, \mathbf{B}) .

2.5 Matrix Analysis

In the derivation and analysis of neural network-based PCA algorithm and its extensions, besides SVD, ED, etc., matrix gradient and matrix differential are also very necessary analysis tools. In this section, we will introduce some important results and properties of matrix gradient and matrix differential.

2.5.1 Differential and Integral of Matrix with Respect to Scalar

If $\mathbf{A}(t) = \{a_{ij}(t)\}_{m \times n}$ is a real matrix function of scalar t , then its differential and integral are, respectively, defined as

$$\begin{cases} \frac{d}{dt} \mathbf{A}(t) = \left\{ \frac{d}{dt} a_{ij}(t) \right\}_{m \times n} \\ \int \mathbf{A}(t) dt = \left\{ \int a_{ij}(t) dt \right\}_{m \times n} \end{cases}.$$

If $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are, respectively, $m \times n$ and $n \times r$ matrices, then

$$\frac{d}{dt} [\mathbf{A}(t) \mathbf{B}(t)] = \left[\frac{d\mathbf{A}(t)}{dt} \right] \mathbf{B}(t) + \mathbf{A}(t) \left[\frac{d\mathbf{B}(t)}{dt} \right].$$

If $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are both $m \times n$ matrices, then

$$\frac{d}{dt}[\mathbf{A}(t) + \mathbf{B}(t)] = \frac{d\mathbf{A}(t)}{dt} + \frac{d\mathbf{B}(t)}{dt}.$$

If $\mathbf{A}(t)$ is a rank- n invertible square matrix, then

$$\frac{d\mathbf{A}^{-1}(t)}{dt} = -\mathbf{A}^{-1}(t) \frac{d\mathbf{A}(t)}{dt} \mathbf{A}^{-1}(t).$$

2.5.2 Gradient of Real Function with Respect to Real Vector

Define gradient operator $\nabla_{\mathbf{x}}$ of an $n \times 1$ vector \mathbf{x} as

$$\nabla_{\mathbf{x}} = \left[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \dots, \frac{\partial}{\partial x_n} \right]^T = \frac{\partial}{\partial \mathbf{x}},$$

Then the gradient of a real scalar quantity function $f(\mathbf{x})$ with respect to \mathbf{x} is a $n \times 1$ column vector, which is defined as

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right]^T = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}.$$

The negative direction of the gradient direction is called as the gradient flow of variable \mathbf{x} , written as

$$\dot{\mathbf{x}} = -\nabla_{\mathbf{x}} f(\mathbf{x}).$$

The gradient of m -dimensional row vector function $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})]$ with respect to the $n \times 1$ real vector \mathbf{x} is an $n \times m$ matrix, defined as

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \frac{\partial f_m(\mathbf{x})}{\partial x_2} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \frac{\partial f_2(\mathbf{x})}{\partial x_n} & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} = \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}).$$

Some properties of gradient operations can be summarized as follows:

- ① If $f(\mathbf{x}) = c$ is a constant, then gradient $\frac{\partial c}{\partial \mathbf{x}} = \mathbf{0}$.
- ② Linear principle: If $f(\mathbf{x})$ and $g(\mathbf{x})$ are real functions of vector \mathbf{x} , and c_1 and c_2 are real constants, then

$$\frac{\partial[c_1 f(\mathbf{x}) + c_2 g(\mathbf{x})]}{\partial \mathbf{x}} = c_1 \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + c_2 \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}.$$

- ③ Product principle: If $f(\mathbf{x})$ and $g(\mathbf{x})$ are real functions of vector \mathbf{x} , then

$$\frac{\partial f(\mathbf{x})g(\mathbf{x})}{\partial \mathbf{x}} = g(\mathbf{x}) \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + f(\mathbf{x}) \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}.$$

- ④ Quotient principle: If $g(\mathbf{x}) \neq 0$, then

$$\frac{\partial f(\mathbf{x})/g(\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{g^2(\mathbf{x})} \left[g(\mathbf{x}) \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} - f(\mathbf{x}) \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} \right].$$

- ⑤ Chain principle: If $\mathbf{y}(\mathbf{x})$ is a vector-valued function of \mathbf{x} , then

$$\frac{\partial f(\mathbf{y}(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}^T(\mathbf{x})}{\partial \mathbf{x}} \frac{\partial f(\mathbf{y})}{\partial \mathbf{y}},$$

where $\frac{\partial \mathbf{y}^T(\mathbf{x})}{\partial \mathbf{x}}$ is an $n \times n$ matrix.

- ⑥ If \mathbf{a} is an $n \times 1$ constant vector, then

$$\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}, \quad \frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$$

- ⑦ If \mathbf{A} and \mathbf{y} are both independent of \mathbf{x} , then

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A} \mathbf{y}, \quad \frac{\partial \mathbf{y}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T \mathbf{y}.$$

- ⑧ If \mathbf{A} is a matrix independent of \mathbf{x} , then

$$\frac{\partial \mathbf{x}^T \mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}, \quad \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A} \mathbf{x} + \mathbf{A}^T \mathbf{x} = (\mathbf{A} + \mathbf{A}^T) \mathbf{x}.$$

Especially, if \mathbf{A} is a symmetric matrix, then $\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A} \mathbf{x}$.

2.5.3 Gradient Matrix of Real Function

The gradient of a real function $f(\mathbf{A})$ with respect to an $m \times n$ real matrix \mathbf{A} is an $m \times n$ matrix, called as gradient matrix, defined as

$$\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} = \begin{bmatrix} \frac{\partial f(\mathbf{A})}{\partial A_{11}} & \frac{\partial f(\mathbf{A})}{\partial A_{12}} & \dots & \frac{\partial f(\mathbf{A})}{\partial A_{1n}} \\ \frac{\partial f(\mathbf{A})}{\partial A_{21}} & \frac{\partial f(\mathbf{A})}{\partial A_{22}} & \dots & \frac{\partial f(\mathbf{A})}{\partial A_{2n}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f(\mathbf{A})}{\partial A_{m1}} & \frac{\partial f(\mathbf{A})}{\partial A_{m2}} & \dots & \frac{\partial f(\mathbf{A})}{\partial A_{mn}} \end{bmatrix} = \nabla_{\mathbf{A}} f(\mathbf{A}),$$

where A_{ij} is the element of matrix \mathbf{A} on its i th row and j th column.

Some properties of the gradient of a real function with respect to a matrix can be summarized as follows:

- ① If $f(\mathbf{A}) = c$ is a constant, where \mathbf{A} is an $m \times n$ matrix, then $\frac{\partial c}{\partial \mathbf{A}} = \mathbf{O}_{m \times n}$.
- ② Linear principle: If $f(\mathbf{A})$ and $g(\mathbf{A})$ are real functions of matrix \mathbf{A} , and c_1 and c_2 are real constants, then

$$\frac{\partial [c_1 f(\mathbf{A}) + c_2 g(\mathbf{A})]}{\partial \mathbf{A}} = c_1 \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} + c_2 \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}.$$

- ③ Product principle: If $f(\mathbf{A})$ and $g(\mathbf{A})$ are real functions of matrix \mathbf{A} , then

$$\frac{\partial f(\mathbf{A})g(\mathbf{A})}{\partial \mathbf{A}} = g(\mathbf{A}) \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} + f(\mathbf{A}) \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}}.$$

- ④ Quotient principle: If $g(\mathbf{A}) \neq 0$, then

$$\frac{\partial f(\mathbf{A})/g(\mathbf{A})}{\partial \mathbf{A}} = \frac{1}{g^2(\mathbf{A})} \left[g(\mathbf{A}) \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} - f(\mathbf{A}) \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}} \right].$$

- ⑤ Chain principle: Let \mathbf{A} be an $m \times n$ matrix, and $y = f(\mathbf{A})$ and $g(y)$ are real functions of matrix \mathbf{A} and scalar y , respectively. Then

$$\frac{\partial g(f(\mathbf{A}))}{\partial \mathbf{A}} = \frac{dg(y)}{dy} \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}}.$$

- ⑥ If $\mathbf{A} \in \Re^{m \times n}$, $\mathbf{x} \in \Re^{m \times 1}$, $\mathbf{y} \in \Re^{n \times 1}$, then

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial \mathbf{A}} = \mathbf{A} \mathbf{y}^T.$$

- ⑦ If $\mathbf{A} \in \Re^{n \times n}$ is nonsingular $\mathbf{x} \in \Re^{n \times 1}$, $\mathbf{y} \in \Re^{n \times 1}$, then

$$\frac{\partial \mathbf{x}^T \mathbf{A}^{-1} \mathbf{y}}{\partial \mathbf{A}} = -\mathbf{A}^{-T} \mathbf{A} \mathbf{y}^T \mathbf{A}^{-T}.$$

⑧ If $A \in \mathbb{R}^{m \times n}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n \times 1}$, then

$$\frac{\partial \mathbf{x}^T A^T A \mathbf{y}}{\partial A} = A (\mathbf{x} \mathbf{y}^T + \mathbf{y} \mathbf{x}^T).$$

⑨ If $A \in \mathbb{R}^{m \times n}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m \times 1}$, then

$$\frac{\partial \mathbf{x}^T A A^T \mathbf{y}}{\partial A} = (\mathbf{x} \mathbf{y}^T + \mathbf{y} \mathbf{x}^T) A.$$

2.5.4 Gradient Matrix of Trace Function

Here, we summarize some properties of gradient matrix of trace functions.

①–③ are gradient matrices of the trace of a single matrix.

① If W is an $m \times m$ matrix, then

$$\frac{\partial \text{tr}(W)}{\partial W} = I_m.$$

② If an $m \times m$ matrix W is invertible, then

$$\frac{\partial \text{tr}(W^{-1})}{\partial W} = -(W^{-2})^T.$$

③ For the outer product of two vectors, it holds that

$$\frac{\partial \text{tr}(\mathbf{x} \mathbf{y}^T)}{\partial \mathbf{x}} = \frac{\partial \text{tr}(\mathbf{y} \mathbf{x}^T)}{\partial \mathbf{x}} = \mathbf{y}.$$

④–⑦ are gradient matrices of the trace of the product of two matrices.

④ If $W \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{n \times m}$, then

$$\frac{\partial \text{tr}(WA)}{\partial W} = \frac{\partial \text{tr}(AW)}{\partial W} = A^T.$$

⑤ If $W \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{m \times n}$, then

$$\frac{\partial \text{tr}(W^T A)}{\partial W} = \frac{\partial \text{tr}(A W^T)}{\partial W} = A.$$

⑥ If $\mathbf{W} \in \mathbb{R}^{m \times n}$, then

$$\frac{\partial \text{tr}(\mathbf{W}\mathbf{W}^T)}{\partial \mathbf{W}} = \frac{\partial \text{tr}(\mathbf{W}^T\mathbf{W})}{\partial \mathbf{W}} = 2\mathbf{W}.$$

⑦ If $\mathbf{W} \in \mathbb{R}^{m \times n}$, then

$$\frac{\partial \text{tr}(\mathbf{W}^2)}{\partial \mathbf{W}} = \frac{\partial \text{tr}(\mathbf{W}\mathbf{W})}{\partial \mathbf{W}} = 2\mathbf{W}^T.$$

⑧ If $\mathbf{W}, \mathbf{A} \in \mathbb{R}^{m \times m}$ and \mathbf{W} is nonsingular, then

$$\frac{\partial \text{tr}(\mathbf{A}\mathbf{W}^{-1})}{\partial \mathbf{W}} = -(\mathbf{W}^{-1}\mathbf{A}\mathbf{W}^{-1})^T.$$

⑨–⑪ are gradient matrices of the trace of the product of three matrices.

⑨ If $\mathbf{W} \in \mathbb{R}^{m \times n}$, $\mathbf{A} \in \mathbb{R}^{m \times m}$, then

$$\frac{\partial \text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})}{\partial \mathbf{W}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{W}.$$

Especially, if \mathbf{A} is a symmetric matrix, then $\frac{\partial \text{tr}(\mathbf{W}^T\mathbf{A}\mathbf{W})}{\partial \mathbf{W}} = 2\mathbf{A}\mathbf{W}$

⑩ If $\mathbf{W} \in \mathbb{R}^{m \times n}$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, then

$$\frac{\partial \text{tr}(\mathbf{W}\mathbf{A}\mathbf{W}^T)}{\partial \mathbf{W}} = \mathbf{W}(\mathbf{A} + \mathbf{A}^T).$$

Especially, if \mathbf{A} is a symmetric matrix, then $\frac{\partial \text{tr}(\mathbf{W}\mathbf{A}\mathbf{W}^T)}{\partial \mathbf{W}} = 2\mathbf{W}\mathbf{A}$

⑪ If $\mathbf{W}, \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times m}$ and \mathbf{W} is nonsingular, then

$$\frac{\partial \text{tr}(\mathbf{A}\mathbf{W}^{-1}\mathbf{B})}{\partial \mathbf{W}} = -(\mathbf{W}^{-1}\mathbf{B}\mathbf{A}\mathbf{W}^{-1})^T.$$

2.5.5 Gradient Matrix of Determinant

Some properties of the gradient of the determinant of a matrix can be summarized as follows:

- ① Gradient of the determinant of a single nonsingular matrix

$$\begin{aligned}\frac{\partial |\mathbf{W}|}{\partial \mathbf{W}} &= |\mathbf{W}|(\mathbf{W}^{-1})^T = (\mathbf{W}^\#)^T \\ \frac{\partial |\mathbf{W}^{-1}|}{\partial \mathbf{W}} &= -|\mathbf{W}|^{-1}(\mathbf{W}^{-1})^T,\end{aligned}$$

where $\mathbf{W}^\#$ is the adjoint matrix \mathbf{A} .

- ② Gradient of the logarithm of a determinant

$$\frac{\partial}{\partial \mathbf{W}} \log |\mathbf{W}| = \frac{1}{|\mathbf{W}|} \frac{\partial |\mathbf{W}|}{\partial \mathbf{W}},$$

\mathbf{W} is nonsingular.

$$\frac{\partial}{\partial \mathbf{W}} \log |\mathbf{W}| = (\mathbf{W}^{-1})^T,$$

the elements are independent to each other.

$$\frac{\partial}{\partial \mathbf{W}} \log |\mathbf{W}| = 2\mathbf{W}^{-1} - \text{diag}(\mathbf{W}^{-1}),$$

\mathbf{W} is symmetric matrix.

- ③ Gradient of the determinant of a two-matrix product

$$\frac{\partial |\mathbf{W}\mathbf{W}^T|}{\partial \mathbf{W}} = 2|\mathbf{W}\mathbf{W}^T|(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W}, \quad \text{rank}(\mathbf{W}_{m \times n}) = m.$$

$$\frac{\partial |\mathbf{W}\mathbf{W}^T|}{\partial \mathbf{W}} = 2|\mathbf{W}^T\mathbf{W}|\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}, \quad \text{rank}(\mathbf{W}_{m \times n}) = n.$$

$$\frac{\partial |\mathbf{W}^2|}{\partial \mathbf{W}} = 2|\mathbf{W}|^2(\mathbf{W}^{-1})^T, \quad \text{rank}(\mathbf{W}_{m \times m}) = m.$$

- ④ Gradient of the determinant of a three-matrix product

$$\frac{\partial |\mathbf{A}\mathbf{W}\mathbf{B}|}{\partial \mathbf{W}} = |\mathbf{A}\mathbf{W}\mathbf{B}|\mathbf{A}^T(\mathbf{B}^T\mathbf{W}^T\mathbf{A}^T)^{-1}\mathbf{B}^T.$$

$$\frac{\partial |\mathbf{W}^T\mathbf{A}\mathbf{W}|}{\partial \mathbf{W}} = 2\mathbf{A}\mathbf{W}(\mathbf{W}^T\mathbf{A}\mathbf{W})^{-1}, \quad |\mathbf{W}^T\mathbf{A}\mathbf{W}| > 0.$$

$$\frac{\partial |\mathbf{W}\mathbf{A}\mathbf{W}^T|}{\partial \mathbf{W}} = \left[(\mathbf{W}\mathbf{A}\mathbf{W}^T)^{-1} \right]^T \mathbf{W}(\mathbf{A}^T + \mathbf{A}).$$

2.5.6 Hessian Matrix

The Hessian matrix is defined as

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \frac{\partial}{\partial \mathbf{x}^T} \left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right] = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$$

and it can also be written as the gradient of gradient, i.e., $\nabla_x^2 f(\mathbf{x}) = \nabla_x (\nabla_x f(\mathbf{x}))$. Here are some properties of Hessian matrix.

- ① For an $n \times 1$ constant vector \mathbf{a} , it holds that

$$\frac{\partial^2 \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x} \partial \mathbf{x}^T} = \mathbf{O}_{n \times n}.$$

- ② If \mathbf{A} is an $n \times n$ matrix, then

$$\frac{\partial^2 \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x} \partial \mathbf{x}^T} = \mathbf{A} + \mathbf{A}^T.$$

- ③ If \mathbf{x} is an $n \times 1$ vector, \mathbf{a} is an $m \times 1$ constant vector, \mathbf{A} and \mathbf{B} , respectively, are $m \times n$ and $m \times m$ constant matrices, and \mathbf{B} is symmetric, then

$$\frac{\partial^2 (\mathbf{a} - \mathbf{A} \mathbf{x})^T \mathbf{B} (\mathbf{a} - \mathbf{A} \mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = 2\mathbf{A}^T \mathbf{B} \mathbf{A}.$$

2.6 Summary

The singular value decomposition, eigenvalue decomposition, Rayleigh quotient, and gradient and differentials of a matrix have been reviewed in a tutorial style in this chapter. The materials presented in this chapter are useful for the understanding of latter chapters, particularly for the chapters except 3 and 6.

References

1. Beltrami, E. (1873). Sulle funzioni bilineari, *Giornale di Matematiche ad Uso studenti Delle Uniersita.* 11, 98–106. (An English translation by D Boley is available as University of Minnesota, Department of Computer Science), Technical Report 90–37, 1990.
2. Jordan, C. (1874). Memoire sur les formes bilineaires. *Journal of Mathematical Pures Application Eeuxieme Series*, 19, 35–54.
3. Autonne, L. (1902). Sur les groupes lineaires, reelles et orthogonaus. *Bulletin Social Math France*, 30, 121–134.
4. Eckart, C., & Young, G. (1939). A principal axis transformation for non-Hermitian matrices. *Null American Mathematics Society*, 45(2), 118–121.
5. Klema, V. C., & Laub, A. J. (1980). The singular value decomposition: Its computation and some application. *IEEE Transactions on Automatic Control*, 25(2), 164–176.
6. Zhang, X. D. (2004). *Matrix analysis and application*. Tsinghua University Press.
7. Jennings, A., & McKeown, J. J. (1992). *Matrix computations*. New York: Wiley.
8. Saad, Y. (1992). *Numerical methods for large eigenvalue problem*. New York: Machester University Press.
9. Dramc, Z. (1998). A tangent algorithm for computing the generalized singular value decomposition. *SIAM Journal of Numerical Analysis*, 35(5), 1804–1832.
10. Nour-Omid, B., Parlett, B. N., Ericsson, T., & Jensen, P. S. (1987). How to implement the spectral transformation. *Mathematics Computation*, 48(178), 663–673.
11. Yang, X., Sarkar, T. K., Yang, X., & Arvas, E. (1989). A survey of conjugate gradient algorithms for solution of extreme Eigen-problems of a symmetric matrix. *IEEE Transactions on Acoustic Speech and Signal Processing*, 37(10), 1550–1556.
12. Rayleigh, L. (1937). *The theory of sound* (2nd ed.). New York: Macmillian.
13. Parlett, B. N. (1974). The Rayleigh quotient iteration and some genelarizations of no normal matrices. *Mathematics of Computation*, 28(127), 679–693.
14. Parlett, B. N. (1980). *The symmetric Eigenvalue problem*. Englewood Cliffs, NJ: Prentice-Hall.
15. Chatelin, F. (1993). *Eigenvalues of matrices*. New York: Wiley.
16. Helmke, U., & Moore, J. B. (1994). *Optimization and dynamical systems*. London, UK: Springer-Verlag.
17. Golub, G. H., & Van Loan, C. F. (1989). *Matrix computation* (2nd ed.). Baltimore: The John Hopkins University Press.
18. Cirrincione, G., Cirrincione, M., Herault, J., & Van Huffel, S. (2002). The MCA EXIN neuron for the minor component analysis. *IEEE Transactions on Neural Networks*, 13(1), 160–187.
19. Lancaster, P., & Tismenetsky, M. (1985). *The theory of Matrices with Applications* (2nd ed.). New York: Academic.
20. Shavitt, I., Bender, C. F., Pipano, A., & Hosteny, R. P. (1973). The iterative calculation of several of the lowest or highest eigenvalues and corresponding eigenvectors of very large symmetric matrices. *Journal of Computation Physics*, 11(1), 90–108.
21. Chen, H., Sarkar, T. K., Brule, J., & Dianat, S. A. (1986). Adaptive spectral estimation by the conjugate gradient method. *IEEE Transactions on Acoustic Speech and Signal Processing*, 34(2), 271–284.
22. Huang, L. (1984). *The linear Algebra in system and control theory*. Beijing: Science Press.
23. Pease, M. C. (1965). *Methods of matrix algebra*. New York: Academic Press.
24. Magnus, J. R., & Neudecker, H. (1999). *Matrix differential calculus with application in statistics and econometrics, revised*. Chichester: Wiley.
25. Lutkepohl, H. (1996). *Handbook of matrices*. New York: Wiley.
26. Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York: Wiley.

Principal Component Analysis Networks and Algorithms

Kong, X.; Hu, C.; Duan, Z.

2017, XXII, 323 p. 86 illus., 41 illus. in color., Hardcover

ISBN: 978-981-10-2913-4