Notes on Statistical and Machine Learning

Random Vectors and Matrices

Chapter: *3* **Prepared by:** *Chenxi Zhou*

This note is prepared based on *Chapter 3, Random Vectors and Matrices* in Izenman (2009).

I. Vector and Matrices

1. **Orthogonal and Idempotent Matrices:** An $n \times n$ matrix $\mathbf{A}$ is said to be *orthogonal* if $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_n$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix, and is *idempotent* if $\mathbf{A}^\top \mathbf{A} = \mathbf{A}$.

2. **Projection Matrix:** An $n \times n$ matrix $\mathbf{P}$ is said to be a *projection matrix* if and only if $\mathbf{P}$ is symmetric and idempotent.

   If $\mathbf{P}$ is both projection and orthogonal, then $\mathbf{P}$ is said to be an orthogonal projector.

3. **Proposition:** If $\mathbf{P}$ is a projection matrix and define $\mathbf{Q} = \mathbf{I} - \mathbf{P}$, then $\mathbf{Q}$ is also a projection matrix.

   Proof. To show that $\mathbf{Q}$ is a projection matrix, we need to show $\mathbf{Q}$ is both symmetric and idempotent:

   - *Symmetry:* $\mathbf{Q}^\top = (\mathbf{I} - \mathbf{P})^\top = \mathbf{I}^\top - \mathbf{P}^\top = \mathbf{I} - \mathbf{P} = \mathbf{Q}$;
   - *Idempotence:* $\mathbf{Q}^2 = \mathbf{Q}\mathbf{Q} = (\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{P}) = \mathbf{I} - \mathbf{P} - \mathbf{P} + \mathbf{P}^2 = \mathbf{I} - \mathbf{P} - \mathbf{P} + \mathbf{P} = \mathbf{I} - \mathbf{P} = \mathbf{Q}$, where we use the idempotence of $\mathbf{P}$ in the third equality.

   ■

4. **Trace:** The *trace* of an $n \times n$ matrix $\mathbf{A}$ is defined to be

   $$\mathrm{trace}(\mathbf{A}) = \sum_{i=1}^{n} A_{i,i},$$

   where $A_{i,i}$ denotes the (i, i)-entry of $\mathbf{A}$.

5. **Properties of Trace:**

   - Let $\mathbf{A}$ and $\mathbf{B}$ both be $n \times n$ square matrices. Then, $\mathrm{trace}(\mathbf{A} + \mathbf{B}) = \mathrm{trace}(\mathbf{A}) + \mathrm{trace}(\mathbf{B})$;
   - Let $\mathbf{A}$ be a $n \times m$ matrix and $\mathbf{B}$ be a $m \times n$ matrix. Then, $\mathrm{trace}(\mathbf{AB}) = \mathrm{trace}(\mathbf{BA})$.

6. **Minor:** Let $\mathbf{A}$ be an $m \times n$ matrix. The minor $\mathbf{M}_{i,j}$ of element $A_{i,j}$ is the $(m-1) \times (n-1)$ matrix formed by deleting the i-th row and j-th column from $\mathbf{A}$.

7. **Cofactor and Determinant:** Let $\mathbf{A}$ be an $n \times n$ matrix. The *cofactor* of $A_{i,j}$ is $C_{i,j} = (-1)^{i+j}|\mathbf{M}_{i,j}|$, where $|\mathbf{M}|$ is the determinant of the matrix $\mathbf{M}$. One way of defining $|\mathbf{A}|$ is by using *Laplace's formula*:

$$|\mathbf{A}| = \sum_{i=1}^{n} A_{i,j} C_{i,j},$$

where we expand along the i-th row.

8. **Some Properties of Determinant:**

   - $|\mathbf{A}^\top| = |\mathbf{A}|$;
   - If a is a scalar and $\mathbf{A}$ is a $n \times n$ matrix, then $|a\mathbf{A}| = a^n \cdot |A|$.

9. **Singular and Nonsingular Matrices:** The $n \times n$ matrix $\mathbf{A}$ is said to be *singular* if $|\mathbf{A}| = 0$, and is *nonsingular* otherwise.

10. **Matrix Decomposition:**

    - *LR Decomposition:* $\mathbf{A} = \mathbf{LR}$, where $\mathbf{L}$ is a lower-triangular matrix and $\mathbf{R}$ is an upper-triangular matrix;
    - *Cholesky Decomposition:* Let $\mathbf{A}$ be a symmetric positive definite matrix. Then, we can write $\mathbf{A} = \mathbf{LL}^\top$, where $\mathbf{L}$ is a lower-triangular matrix;
    - *QR-Decomposition:* $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is upper-triangular.

11. **Determinant of a Partitioned Matrix:** Let

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$$

be a partitioned matrix, where $\mathbf{A}$ and $\mathbf{D}$ are both square and nonsingular. Then, the determinant of $\mathbf{\Sigma}$ can be expressed as

$$|\mathbf{\Sigma}| = |\mathbf{A}| \cdot |\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{D}| \cdot |\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}|.$$

12. **Rank:** The *rank* of a matrix $\mathbf{A}$, denoted $\text{rank}(A)$, is the size of the largest sub-matrix of $\mathbf{A}$ that has a nonzero determinant; it is also the number of linearly independent rows/columns of $\mathbf{A}$.

13. **Properties of Rank:**

    (a) $\text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{A})$ if $|\mathbf{B}| \neq 0$;

    (b) $\text{rank}(\mathbf{AB}) \leq \min\{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\}$.

14. **Inverse:**

(a) *Definition:* If $\mathbf{A}$ is an $n \times n$ square nonsingular matrix, then a unique $n \times n$ inverse matrix $\mathbf{A}^{-1}$ exists such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$.

(b) *Properties:*

- If $\mathbf{A}$ is *orthogonal*, then $\mathbf{A}^{-1} = \mathbf{A}^\top$;
- $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, and $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$;
-
$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1},$$
where $\mathbf{A}$ and $\mathbf{D}$ are $n \times n$ and $m \times m$ nonsingular matrices, respectively;
- If $\mathbf{A}$ is $n \times n$ and $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^n$ are vectors, then, a special case of the previous result is
$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{(\mathbf{A}^{-1}\mathbf{u})(\mathbf{v}^\top\mathbf{A}^{-1})}{1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}},$$
reducing the problem of inverting $\mathbf{A} + \mathbf{u}\mathbf{v}^\top$ to the one of just inverting $\mathbf{A}$;
- If $\mathbf{A}$ and $\mathbf{D}$ are symmetric and $\mathbf{A}$ is nonsingular, then,
$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}^\top & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}^\top & \mathbf{E}^{-1} \end{pmatrix},$$
where $\mathbf{E} := \mathbf{D} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}$ is nonsingular and $\mathbf{F} := \mathbf{A}^{-1}\mathbf{B}$.

15. Quadratic Form: If $\mathbf{A}$ is an $n \times n$-matrix and $\mathbf{x} \in \mathbb{R}^n$ is a vector, then a quadratic form is

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}x_ix_j,$$

where $A_{i,j}$ is the (i,j)-entry of $\mathbf{A}$ and $\mathbf{x} = (x_1, x_2, \cdots, x_n)^\top \in \mathbb{R}^n$. An $n \times n$-matrix $\mathbf{A}$ is

(a) *positive-definite* if, for any n-vector $\mathbf{x} \neq \mathbf{0}_n$, the quadratic form $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$, and

(b) *nonnegative-definite* or *positive-semidefinite* if $\mathbf{x}^\top\mathbf{A}\mathbf{x} \geq 0$.

16. Vectoring Operation: Let $\mathbf{A}$ be an $m \times n$ matrix and the vectoring operator $\mathrm{vec}(\mathbf{A})$ denotes the $mn \times 1$-column vector by placing the columns of $\mathbf{A}$ under one another successively.

17. Kronecker Product: Let $\mathbf{A}$ be an $m \times n$-matrix and $\mathbf{B}$ be an $s \times t$-matrix. Then, the *(left) Kronecker product* of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \otimes \mathbf{B}$ is the $ms \times nt$ block matrix

$$\mathbf{A} \otimes \mathbf{B} = [\mathbf{A}B_{j,k}] = \begin{pmatrix} \mathbf{A}B_{1,1} & \cdots & \mathbf{A}B_{1,t} \\ \vdots & \ddots & \vdots \\ \mathbf{A}B_{s,1} & \cdots & \mathbf{A}B_{s,t} \end{pmatrix}. \tag{1}$$

The *right Kronecker product* of $\mathbf{A}$ and $\mathbf{B}$ is defined to be $[A_{i,j}\mathbf{B}]$.

18. **Properties of Kronecker Product:**
    - $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$;
    - $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$;
    - $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = (\mathbf{A} \otimes \mathbf{C}) + (\mathbf{B} \otimes \mathbf{C})$;
    - $(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{A}^\top \otimes \mathbf{B}^\top$;
    - $\mathrm{trace}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{trace}(\mathbf{A}) \cdot \mathrm{trace}(\mathbf{B})$;
    - $\mathrm{rank}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{rank}(\mathbf{A}) \cdot \mathrm{rank}(\mathbf{B})$;
    - If $\mathbf{A}$ is of size $n \times n$ and $\mathbf{B}$ is of size $m \times m$, then $|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^m \cdot |\mathbf{B}|^n$;
    - If $\mathbf{A}$ is of size $m \times n$ and $\mathbf{B}$ is of size $s \times t$, then, $\mathbf{A} \otimes \mathbf{B} = (\mathbf{A} \otimes \mathbf{I}_s)(\mathbf{I}_n \otimes \mathbf{B})$;
    - If $\mathbf{A}$ and $\mathbf{B}$ are square and nonsingular, then $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$;
    - $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{A} \otimes \mathbf{C}^\top)\, \mathrm{vec}(\mathbf{B})$.

19. **Outer Product:** The *outer product* of $\mathbf{v} \in \mathbb{R}^n$ with itself is the $n \times n$-matrix $\mathbf{v}\mathbf{v}^\top$, which has rank 1.

20. **Characteristic Polynomial, Eigenvalues and Eigenvectors:** If $\mathbf{A}$ is a matrix of size $n \times n$, then $|\mathbf{A} - \lambda \mathbf{I}_n|$, called the *characteristic polynomial*, is a polynomial of order n in λ.

    The equation $|\mathbf{A} - \lambda \mathbf{I}_n| = 0$ will have n (possibly complex-valued, not necessarily distinct) roots denoted by $\lambda_i := \lambda_i(\mathbf{A})$ for $i = 1, 2, \cdots, n$. The root λ_i is called an *eigenvalue* of $\mathbf{A}$, and the set $\{\lambda_i\}_{i=1}^n$ is called the *spectrum* of $\mathbf{A}$.

    Associated with λ_i, there is a nonzero vector $\mathbf{v}_i := \mathbf{v}_i(\mathbf{A}) \in \mathbb{R}^n$ (not all of whose entries of zero) such that $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$. The vector $\mathbf{v}_i$ is called an *eigenvector* associated with λ_i.

    Remark. Eigenvalues of a positive-definite matrix are all positive, and eigenvalues of a nonnegative-definite matrix are all nonnegative.

21. **Properties of Eigenvalues and Eigenvectors:** Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric real matrix.
    - All eigenvalues of $\mathbf{A}$ are real;
    - Eigenvectors $\mathbf{v}_i$ and $\mathbf{v}_j$ associated with distinct eigenvalues $\lambda_i \neq \lambda_j$ are *orthogonal*;
    - If $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n) \in \mathbb{R}^{n \times n}$, then

      $$\mathbf{AV} = \mathbf{V}\boldsymbol{\Lambda},$$

      where $\boldsymbol{\Lambda} := \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) \in \mathbb{R}^{n \times n}$ is a matrix with the eigenvalues along the diagonal and zeroes elsewhere, and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_n$;

- *Spectral Theorem:* One can write the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ as a weighted average of rank-1 matrices, i.e.,

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^\top,$$

  where $\mathbf{I}_n = \sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i^\top$ and the weights, $\lambda_1, \cdots, \lambda_n$, are the eigenvalues of $\mathbf{A}$;
- The rank of $\mathbf{A}$ is the number of nonzero eigenvalues;
- The trace of $\mathbf{A}$ is equal to the sum of all eigenvalues, i.e., $\text{trace}(\mathbf{A}) = \sum_{i=1}^n \lambda_i(\mathbf{A})$;
- The determinant of $\mathbf{A}$ is equal to the product of all eigenvalues, i.e., $|\mathbf{A}| = \prod_{i=1}^n \lambda_i(\mathbf{A})$.

Remark. Some of the results above also hold for a general square matrix (not necessarily symmetric).

22. Functions of Matrices: Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix and let $\phi : \mathbb{R}^n \to \mathbb{R}^n$ be a function. Then,

$$\phi(\mathbf{A}) = \sum_{i=1}^n \phi(\lambda_i)\mathbf{v}_i\mathbf{v}_i^\top,$$

where λ_i is the i-th eigenvalue $\mathbf{A}$, and $\mathbf{v}_i$ is the corresponding eigenvector.

Examples:

- Suppose $\mathbf{A}$ is nonsingular. Then,

$$\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^\top = \sum_{i=1}^n \lambda_i^{-1}\mathbf{v}_i\mathbf{v}_i^\top;$$

- Suppose $\mathbf{A}$ is nonnegative-definite. Then,

$$\mathbf{A}^{1/2} = \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top = \sum_{i=1}^n \sqrt{\lambda_i}\mathbf{v}_i\mathbf{v}_i^\top;$$

- Suppose $\mathbf{A}$ is positive-definite. Then,

$$\log(\mathbf{A}) = \sum_{i=1}^n \log(\lambda_i)\mathbf{v}_i\mathbf{v}_i^\top.$$

23. Proposition: If $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $m \le n$, then

$$\lambda_i(\mathbf{A}^\top\mathbf{A}) = \lambda_i(\mathbf{A}\mathbf{A}^\top), \qquad \text{for all } i = 1, 2, \cdots, m,$$

and $\lambda_i = 0$ for all $i = m+1, m+2, \cdots, n$. Furthermore, for $\lambda_j(\mathbf{A}\mathbf{A}^\top) \neq 0$,

$$\mathbf{v}_j(\mathbf{A}^\top\mathbf{A}) = \sqrt{\lambda_j(\mathbf{A}\mathbf{A}^\top)}\mathbf{A}^\top\mathbf{v}_j(\mathbf{A}\mathbf{A}^\top),$$
$$\mathbf{v}_j(\mathbf{A}\mathbf{A}^\top) = \sqrt{\lambda_j(\mathbf{A}\mathbf{A}^\top)}\mathbf{A}\mathbf{v}_j(\mathbf{A}^\top\mathbf{A}).$$

24. **Singular-Value Decomposition:** The *singular-value decomposition* (SVD) of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, where $m \leq n$, is given by

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Psi}\mathbf{V}^\top = \sum_{i=1}^{m} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^\top. \tag{2}$$

Here,

- $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_m) \in \mathbb{R}^{m\times m}$ and $\mathbf{u}_i = \mathbf{v}_i(\mathbf{A}\mathbf{A}^\top)$ for all $i = 1, 2, \cdots, m$;
- $\mathbf{V} = (\mathbf{v}1, \cdots, \mathbf{v}_n) \in \mathbb{R}^{n\times n}$ and $\mathbf{v}_j = \mathbf{v}_j(\mathbf{A}^\top\mathbf{A})$ for all $j = 1, 2, \cdots, n$;
- $\lambda_i = \lambda_i(\mathbf{A}\mathbf{A}^\top)$ for all $i = 1, 2, \cdots, m$, and

$$\boldsymbol{\Psi} := \begin{pmatrix} \boldsymbol{\Psi}_\sigma & \vdots & \mathbf{0}_{m\times(n-m)} \end{pmatrix}.$$

is a $m \times n$-matrix, and $\boldsymbol{\Psi}_\sigma$ is an $m \times m$ diagonal matrix with the nonnegative *singular values*, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$, of $\mathbf{A}$ along the diagonal, where $\sigma_i = \sqrt{\lambda_i}$ is the square-root of the i-th largest eigenvalue of the $m \times m$-matrix $\mathbf{A}\mathbf{A}^\top$ for all $i = 1, 2, \cdots, m$.

25. **A Direct Consequence of SVD:** Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $m \leq n$. If $\mathrm{rank}(\mathbf{A}) = t$, then there exists matrices $\mathbf{B} \in \mathbb{R}^{m\times t}$ and $\mathbf{C} \in \mathbb{R}^{t\times n}$, both of rank t, such that $\mathbf{A} = \mathbf{B}\mathbf{C}$.

26. **Generalized Inverse:** A *g-inverse* of a $m\times n$-matrix $\mathbf{A}$ is any $n\times m$-matrix, denoted by $\mathbf{A}^-$, such that, for any m-vector $\mathbf{y}$ for which $\mathbf{A}\mathbf{x} = \mathbf{y}$ is a consistent equation, $\mathbf{x} = \mathbf{A}^-\mathbf{y}$ is a solution. We call such an $\mathbf{A}^-$ a *reflexive g-inverse*.

27. **Proposition (Existence of g-Inverse):** $\mathbf{A}^-$ exists if and only if $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$.

28. **Proposition:** A general solution of the consistent equation $\mathbf{A}\mathbf{x} = \mathbf{y}$ is given by

$$\mathbf{x} = \mathbf{A}^-\mathbf{y} + (\mathbf{A}^-\mathbf{A} - \mathbf{I}_n)\,\mathbf{z},$$

where $\mathbf{z} \in \mathbb{R}^n$ is arbitrary.

Remark. The consequence of the preceding proposition is that the g-inverse of a matrix is *not* unique.

Remark. If we let $\mathbf{z} = \mathbf{0}_n$ in (3), the resulting $\mathbf{x} = \mathbf{A}^-\mathbf{y}$ has the minimum norm among all solutions to $\mathbf{A}\mathbf{x} = \mathbf{y}$.

29. **Moore-Penrose Generalized Inverse:** Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a matrix with the SVD $\mathbf{A} = \mathbf{U}\boldsymbol{\Psi}\mathbf{V}^\top$. Then, the unique *Moore-Penrose generalized inverse* of $\mathbf{A}$ is given by

$$\mathbf{A}^\dagger = \mathbf{V}\boldsymbol{\Psi}^\dagger\mathbf{U}^\top,$$

where $\boldsymbol{\Psi}^\dagger$ is a "diagonal" matrix whose diagonal elements are the reciprocals of the *nonzero* elements of $\boldsymbol{\Psi} = \boldsymbol{\Lambda}^{1/2}$, and zeroes otherwise.

30. **Properties of Moore-Penrose Generalized Inverse:**

(a) $\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}$;

(b) $\mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}$;

(c) $(\mathbf{A}\mathbf{A}^{\dagger})^{\top} = \mathbf{A}\mathbf{A}^{\dagger}$;

(d) $(\mathbf{A}^{\dagger}\mathbf{A})^{\top} = \mathbf{A}^{\dagger}\mathbf{A}$.

31. Matrix Norm: The *norm* of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a function $\|\cdot\|$ mapping from $\mathbb{R}^{m\times n}$ to $\mathbb{R}$ satisfying the following conditions:

(a) $\|\mathbf{A}\| \geq 0$;

(b) $\|\mathbf{A}\| = 0$ iff $\mathbf{A} = \mathbf{0}_{m\times n}$;

(c) $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$;

(d) $\|\alpha\mathbf{A}\| = |\alpha| \cdot \|\mathbf{A}\|$.

In the definition above, $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ and $\alpha \in \mathbb{R}$.

32. Examples of Matrix Norms: Let $\mathbf{A} \in \mathbb{R}^{m\times n}$.

(a) *p-norm:*

$$\|\mathbf{A}\|_p = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{i,j}|^p\right)^{1/p};$$

(b) *Frobenius norm:*

$$\|\mathbf{A}\|_F := \sqrt{\operatorname{trace}(\mathbf{A}\mathbf{A}^{\top})} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}A_{i,j}^2\right)^{1/2} = \left(\sum_{i=1}^{m}\lambda_j(\mathbf{A}\mathbf{A}^{\top})\right)^{1/2};$$

(c) *Spectral norm:* Let $m = n$ so that $\mathbf{A}$ is a square matrix, the *spectral norm* is

$$\sqrt{\lambda_1(\mathbf{A}\mathbf{A}^{\top})}.$$

33. Condition Number: The *condition number* of a nonsingular square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is given by

$$\kappa(\mathbf{A}) := \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| = \frac{\sigma_1}{\sigma_n}, \tag{3}$$

which is the ratio of the largest to the smallest nonzero singular value. Here, the norm is taken to be the *spectral norm* and σ_i is the square-root of the i-th largest eigenvalue of the $n \times n$-matrix $\mathbf{A}^{\top}\mathbf{A}$, for all $i = 1, 2, \cdots, n$.

34. Well-conditioned and Ill-conditioned Matrices: The matrix $\mathbf{A}$ is said to be *ill-conditioned* if its singular values are widely spread out, so that $\kappa(\mathbf{A})$ is large, and $\mathbf{A}$ is said to be *well-conditioned* if $\kappa(\mathbf{A})$ is small.

35. **Eckart-Young Theorem:** Let $\mathbf{A}$ and $\mathbf{B}$ are both $(m \times n)$-matrices. Suppose $\mathbf{A}$ is of full rank with $\text{rank}(\mathbf{A}) = \min\{m, n\}$ and $\mathbf{B}$ is of reduced rank with $r_{\mathbf{B}} := \text{rank}(\mathbf{B}) < \min\{m, n\}$. Suppose we want to use $\mathbf{B}$ to approximate $\mathbf{A}$. Then,

$$\lambda_j\big((\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^\top\big) \geq \lambda_{j+r_{\mathbf{B}}}(\mathbf{A}\mathbf{A}^\top), \tag{4}$$

with equality if

$$\mathbf{B} = \sum_{i=1}^{r_{\mathbf{B}}} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^\top,$$

where $\lambda_i = \lambda_i(\mathbf{A}\mathbf{A}^\top)$, $\mathbf{u}_i$ is the eigenvector associated with the i-th largest eigenvalue of $\mathbf{A}\mathbf{A}^\top$ for all $i = 1, 2, \cdots, m$, and $\mathbf{v}_j$ is the eigenvector associated with the j-th largest eigenvalue of $\mathbf{A}^\top\mathbf{A}$, for all $j = 1, 2, \cdots, n$.

Remark. Because the choice of $\mathbf{B}$ provides a simultaneous minimization for *all* eigenvalues λ_i, it follows that the minimum is achieved for different functions of those eigenvalues, e.g., the trace or the determinant of $(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^\top$.

36. **Courant-Fischer Min-Max Theorem:** The i-th largest eigenvalue of a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ can be expressed as

$$\lambda_i(\mathbf{A}) = \inf_{\mathbf{L}} \sup_{\{\mathbf{x} \,|\, \mathbf{L}\mathbf{x} = \mathbf{0}_{i-1}, \mathbf{x} \neq \mathbf{0}_n\}} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}, \tag{5}$$

where the "inf" is taken over $\mathbf{L} \in \mathbb{R}^{(i-1)\times n}$ with rank at most $i - 1$, and the "sup" is the supremum over a nonzero $\mathbf{x} \in \mathbb{R}^n$ that satisfies $\mathbf{L}\mathbf{x} = \mathbf{0}_{i-1}$.

Remark. In (5), the equality is achieved if $\mathbf{L} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{i-1})^\top \in \mathbb{R}^{(i-1)\times n}$ and $\mathbf{x}$ is the eigenvector associated with the i-th largest eigenvalue.

37. **Corollaries of Courant-Fischer Min-Max Theorem:**

(a) The i-th smallest eigenvalue of $\mathbf{A}$ can be written as

$$\lambda_{n-i+1}(\mathbf{A}) = \sup_{\mathbf{L}} \inf_{\{\mathbf{x} \,|\, \mathbf{L}\mathbf{x} = \mathbf{0}_{n-i+1}, \mathbf{x} \neq \mathbf{0}_n\}} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}.$$

(b) The following inequalities hold

$$\lambda_n(\mathbf{A}) \leq \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \leq \lambda_1(\mathbf{A}), \qquad \text{for all } \mathbf{x} \neq \mathbf{0}_n.$$

38. **Hoffman-Wielandt Theorem:** Suppose $\mathbf{A}$ and $\mathbf{B}$ are both symmetric $(n \times n)$-matrices. Suppose $\mathbf{A}$ and $\mathbf{B}$ have eigenvalues $\{\lambda_i(\mathbf{A})\}_{i=1}^n$ and $\{\lambda_i(\mathbf{B})\}_{i=1}^n$, respectively. Then,

$$\sum_{i=1}^n (\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{A}))^2 \leq \text{trace}\big((\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^\top\big). \tag{6}$$

39. **Poincaré Separation Theorem:** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and let $\mathbf{U} \in \mathbb{R}^{n\times m}$ with $m \leq n$ such that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_m$. Then,

$$\lambda_i(\mathbf{U}^\top \mathbf{A}\mathbf{U}) \leq \lambda_i(\mathbf{A}), \qquad \text{for all } i = 1, 2, \cdots, m, \tag{7}$$

with equality being held if the columns of $\mathbf{U}$ are the first m eigenvectors of $\mathbf{A}$.

40. **Matrix Calculus:**

(a) *Jacobian Matrix:* Let $\mathbf{x} = (x_1, \cdots, x_n)^\top \in \mathbb{R}^n$ and

$$\mathbf{y} = (y_1, \cdots, y_m)^\top = \big(f_1(\mathbf{x}), \cdots, f_m(\mathbf{x})\big)^\top = \mathbf{f}(\mathbf{x}) \in \mathbb{R}^m,$$

where $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$. Then, the *partial derivative* of $\mathbf{y}$ with respect to $\mathbf{x}$ is the (mn)-vector

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \left(\frac{\partial y_1}{\partial x_1}, \cdots, \frac{\partial y_m}{\partial x_1}, \cdots, \frac{\partial y_1}{\partial x_n}, \cdots, \frac{\partial y_m}{\partial x_n}\right)^\top.$$

The partial derivative of $\mathbf{y}$ with respect to $\mathbf{x}^\top$ is the $(m \times n)$-matrix

$$\mathbf{J}_{\mathbf{x}}\mathbf{y} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix},$$

called the *Jacobian matrix*.

The Jacobian matrix can be used for linear approximation of a multivariate vector-valued function, i.e.,

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{c}) + [\mathbf{J}_{\mathbf{x}}\mathbf{f}(\mathbf{c})](\mathbf{x} - \mathbf{c}), \qquad \text{for } \mathbf{c} \in \mathbb{R}^n.$$

(b) *Gradient Vector:*

i. If $f : \mathbb{R}^n \to \mathbb{R}$ is a scalar function, then the *gradient vector* is

$$\nabla f(\mathbf{x}) = \frac{\partial y}{\partial \mathbf{x}} = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \cdots, \frac{\partial y}{\partial x_n}\right)^\top = \left(\frac{\partial y}{\partial \mathbf{x}^\top}\right)^\top = (\mathbf{J}_{\mathbf{x}} y)^\top.$$

ii. If $\mathbf{f} : \mathbb{R} \to \mathbb{R}^m$ is a vector function, then the *gradient vector* is

$$\frac{\partial \mathbf{y}}{\partial x} = \left(\frac{\partial y_1}{\partial x}, \frac{\partial y_2}{\partial x}, \cdots, \frac{\partial y_m}{\partial x}\right)^\top.$$

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

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}^\top} = \mathbf{A}, \qquad \frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}^\top} = 2\mathbf{x}, \qquad \frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}^\top} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top).$$

(c) *Derivative of a Matrix:* The *derivative* of an $m \times n$ matrix $\mathbf{A}$ wrt an r-vector $\mathbf{x}$ is the $(mr) \times n$ matrix of derivatives of $\mathbf{A}$ wrt each element of $\mathbf{x}$

$$\frac{\partial \mathbf{A}}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{A}^\top}{\partial x_1}, \cdots, \frac{\partial \mathbf{A}^\top}{\partial x_r} \right)^\top.$$

(d) *Properties of Derivatives of a Matrix:* Let $\mathbf{A}$ and $\mathbf{B}$ be conformable matrices. Then, we have the following

$$\begin{aligned}
\frac{\partial(\alpha\mathbf{A})}{\partial \mathbf{x}} &= \alpha \frac{\partial \mathbf{A}}{\partial \mathbf{x}}, \\
\frac{\partial(\mathbf{A}+\mathbf{B})}{\partial \mathbf{x}} &= \frac{\partial \mathbf{A}}{\partial \mathbf{x}} + \frac{\partial \mathbf{B}}{\partial \mathbf{x}}, \\
\frac{\partial \mathbf{A}\mathbf{B}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{A}}{\partial \mathbf{x}}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial \mathbf{x}}, \\
\frac{\partial \mathbf{A} \otimes \mathbf{B}}{\partial \mathbf{x}} &= \left(\frac{\partial \mathbf{A}}{\partial \mathbf{x}} \otimes \mathbf{B} \right) + \left(\mathbf{A} \otimes \frac{\partial \mathbf{B}}{\partial \mathbf{x}} \right), \\
\frac{\partial \mathbf{A}^{-1}}{\partial \mathbf{x}} &= -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \mathbf{x}}\mathbf{A}^{-1}.
\end{aligned}$$

(e) *Gradient Matrix:* If $y = f(\mathbf{A})$ is a scalar function of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, then the *gradient matrix* is defined to be

$$\frac{\partial y}{\partial \mathbf{A}} = \begin{pmatrix} \frac{\partial y}{\partial A_{1,1}} & \frac{\partial y}{\partial A_{1,2}} & \cdots & \frac{\partial y}{\partial A_{1,n}} \\ \frac{\partial y}{\partial A_{2,1}} & \frac{\partial y}{\partial A_{2,2}} & \cdots & \frac{\partial y}{\partial A_{2,n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial A_{m,1}} & \frac{\partial y}{\partial A_{m,2}} & \cdots & \frac{\partial y}{\partial A_{m,n}} \end{pmatrix}.$$

Examples: Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, then

$$\frac{\partial \operatorname{trace}(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{I}_n, \qquad \text{and} \qquad \frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}| \cdot (\mathbf{A}^\top)^{-1}.$$

(f) *Hessian Matrix:* Let $y = f(\mathbf{x})$ be a scalar function of $\mathbf{x} \in \mathbb{R}^n$. Then, the *Hessian matrix* of y wrt $\mathbf{x}$ is the $n \times n$ matrix

$$\mathbf{H}f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \left(\frac{\partial y}{\partial \mathbf{x}} \right)^\top = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix}$$

Note that $\mathbf{H}f(\mathbf{x}) = \nabla^2_{\mathbf{x}} y = \nabla_{\mathbf{x}} \nabla_{\mathbf{x}} y$ so that the Hessian matrix is the Jacobian of the gradient of f.

The Hessian can be used for a better approximation to a real-valued function f by including a *quadratic* term: for $\mathbf{c} \in \mathbb{R}^n$,

$$f(\mathbf{x}) \approx f(\mathbf{c}) + [\mathbf{J}f(\mathbf{c})](\mathbf{x} - \mathbf{c}) + \frac{1}{2}(\mathbf{x} - \mathbf{c})^\top [\mathbf{H}f(\mathbf{c})](\mathbf{x} - \mathbf{c}). \tag{8}$$

II. Random Vectors

1. **Random Vector:** Suppose we have p random variables, $X_1, \cdots, X_p$, each defined on the real line, and we can write them jointly as a p-dimensional column vector

$$X = (X_1, \cdots, X_p)^\top.$$

2. **Joint Cumulative Distribution Function:** The *joint distribution function* F_X of the random vector X is given by

$$F_X(\mathbf{x}) = F_X(x_1, \cdots, x_p) = \mathbb{P}(X_1 \le x_1, \cdots, X_p \le x_p) = \mathbb{P}(X \le \mathbf{x}),$$

for any $\mathbf{x} = (x_1, x_2, \cdots, x_p)^\top$.

3. **Joint Density Function:** If F_X is absolutely continuous, the *joint density function* f_X of the random vector X is

$$f_X(\mathbf{x}) = f_X(x_1, \cdots, x_p) = \frac{\partial^p F_X(u_1, u_2, \cdots, u_p)}{\partial u_1 \partial u_2 \cdots \partial u_p}\bigg|_{\mathbf{u}=\mathbf{x}},$$

which exists almost everywhere, where $\mathbf{u} = (u_1, u_2, \cdots, u_p)^\top$.

Relationship between Joint Cumulative Distribution Function and Joint Density Function: One can obtain the joint cumulative distribution function F_X and the joint density function f_X by

$$\begin{aligned} F_X(\mathbf{x}) &= F_X(x_1, \cdots, x_p) \\ &= \int_{-\infty}^{x_p} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_X(u_1, u_2, \cdots, u_p) \, \mathrm{d}u_1 \mathrm{d}u_2 \cdots \mathrm{d}u_p. \end{aligned}$$

4. **Marginal Distribution and Density Functions:** Let $(X_1, \cdots, X_k)$, with $k < p$, be a subset of the random vector $X = (X_1, \cdots, X_p)$. The marginal distribution function of this subset is

$$\begin{aligned} F_X(x_1, \cdots, x_k) &= F_X(x_1, \cdots, x_k, \infty, \cdots, \infty) \\ &= \mathbb{P}(X_1 \le x_1, \cdots, X_k \le x_k, X_{k+1} \le \infty, \cdots, X_r \le \infty), \end{aligned}$$

and the marginal density function of the subset is

$$f_X(x_1, \cdots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(u_1, \cdots, u_p) \, \mathrm{d}u_{k+1} \cdots \mathrm{d}u_p.$$

5. **Independence:** The components of a random vector $X \in \mathbb{R}^p$ are said to be *mutually independent* if the joint distribution can be factored into the product of its p marginals, i.e.,

$$F_X(\mathbf{x}) = \prod_{i=1}^{p} F_{X_i}(x_i)$$

where F_{X_i} is the marginal distribution function of X_i for all $i = 1, 2, \cdots, p$. This also means that the joint density function can be factored in the following way under independence,

$$f_X(\mathbf{x}) = \prod_{i=1}^{p} f_{X_i}(x_i).$$

6. **Expectation of a Random Vector:** If $X \in \mathbb{R}^p$ is a random vector, its expected value is the following p-dimensional vector

$$\boldsymbol{\mu}_X = \mathbb{E}[X] = \big(\mathbb{E}[X_1], \cdots, \mathbb{E}[X_p]\big)^\top = (\mu_1, \cdots, \mu_p)^\top \in \mathbb{R}^p.$$

7. **Covariance Matrix:** The $p \times p$ *covariance matrix* of a p-dimensional random vector X is given by

$$\begin{aligned}
\boldsymbol{\Sigma}_{XX} &= \mathrm{Cov}(X, X) \\
&= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] \\
&= \mathbb{E}[(X_1 - \mu_1, \cdots, X_p - \mu_p)(X_1 - \mu_1, \cdots, X_p - \mu_p)^\top] \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_p^2 \end{pmatrix},
\end{aligned}$$

where

$$\sigma_i^2 := \mathrm{Var}(X_i) = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2]$$

is the *variance* of X_i for $i = 1, \cdots, p$ and

$$\sigma_{i,j} := \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$$

is the *covariance* between X_i and X_j for $i, j = 1, \cdots, p$ and $i \neq j$.

8. **Correlation Matrix:** The *correlation matrix* of a p-dimensional random vector X is obtained from the covariance matrix $\boldsymbol{\Sigma}_{XX}$ by dividing the i-th row by σ_i and dividing the j-th column by σ_j, which is given by the following $p \times p$ matrix

$$\mathbf{P}_{XX} = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,p} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p,1} & \rho_{p,2} & \cdots & \rho_{p,p} \end{pmatrix},$$

where

$$\rho_{i,j} = \rho_{j,i} = \begin{cases} \dfrac{\sigma_{i,j}}{\sigma_i \sigma_j}, & \text{if } i \neq j \\ 1, & \text{otherwise} \end{cases}$$

is the *pairwise correlation coefficient* of X_i with X_j for $i, j = 1, \cdots, p$.

Remark. The correlation coefficient $\rho_{i,j}$ lies between -1 and $+1$ and is a measure of association between X_i and X_j:

(a) When $\rho_{i,j} = 0$, we say that X_i and X_j are *uncorrelated*;

(b) When $\rho_{i,j} > 0$, we say that X_i and X_j are *positively correlated*; and

(c) When $\rho_{i,j} < 0$, we say that X_i and X_j are *negatively correlated*.

9. **Stacking Two Random Vectors:** Suppose X and Y are two random vectors, where X is p-dimensional and Y is q-dimensional. Let Z be the random vector of $(p+q)$-dimensional,

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}.$$

Then, the expected value of Z is the $(p+q)$-dimensional vector

$$\boldsymbol{\mu}_Z = \mathbb{E}[Z] = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[Y] \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix},$$

and the covariance matrix of Z is the following partitioned matrix of size $(p+q)\times(p+q)$

$$\begin{aligned} \boldsymbol{\Sigma}_{ZZ} &= \mathbb{E}[(Z - \boldsymbol{\mu}_Z)(Z - \boldsymbol{\mu}_Z)^\top] \\ &= \begin{pmatrix} \mathrm{Cov}(X, X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & \mathrm{Cov}(Y, Y) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}, \end{aligned}$$

where

$$\boldsymbol{\Sigma}_{XY} = \mathrm{Cov}(X, Y) = \mathbb{E}[(X - \boldsymbol{\mu}_X)(Y - \boldsymbol{\mu}_Y)^\top] = \boldsymbol{\Sigma}_{YX}^\top \in \mathbb{R}^{p\times q}.$$

10. **Linearly Related Random Vectors:** If the q-dimensional random vector Y is linearly related to the p-dimensional random vector X in the sense that

$$Y = \mathbf{A}X + \mathbf{b},$$

where $\mathbf{A}$ is a fixed matrix of size $p \times q$ and $\mathbf{b}$ is a q-dimensional fixed vector, then the mean vector and covariance matrix of Y are given by

$$\begin{aligned} \boldsymbol{\mu}_Y &= \mathbf{A}\boldsymbol{\mu}_X + \mathbf{b}, \\ \boldsymbol{\Sigma}_{YY} &= \mathbf{A}\boldsymbol{\Sigma}_{XX}\mathbf{A}^\top, \end{aligned}$$

respectively.

III. Multivariate Gaussian Distribution

1. **Review of a Gaussian Random Variable:** The real-valued univariate random variable X is said to have *Gaussian distribution* with mean μ and variance σ^2, written as $X \sim \text{Normal}(\mu, \sigma^2)$, if its density function is given by

$$f(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \qquad \text{for all } x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$.

2. **Gaussian Random Vector:** The p-dimensional random vector X is said to have the p-variate *Gaussian distribution* with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, which is positive-definite and symmetric, written as $X \sim \text{Normal}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its density function is given by

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^p.$$

3. **Mahalanobis Distance:** The square-root, Δ, of the quadratic form,

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}),$$

is called the *Mahalanobis distance* from $\mathbf{x}$ to μ.

4. **Singular Multivariate Gaussian Distribution:** If $\boldsymbol{\Sigma}$ is *singular*, then, almost surely, the random vector X lives on some hyperplane of *reduced* dimensionality and its density function does *not* exist. In this case, X is said to have a *singular* Gaussian distribution.

5. **Cramer-Wold Theorem:** The distribution of a p-dimensional random vector X is *completely* determined by its one-dimensional linear projections, $\boldsymbol{\alpha}^\top X$, for any vector $\boldsymbol{\alpha} \in \mathbb{R}^p$. More precisely, the random vector X has the multivariate Gaussian distribution if and only if *every* linear function of X has the univariate Gaussian distribution.

6. **Spherical Gaussian Density:** If $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p$, then the multivariate Gaussian density function becomes

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi)^{p/2}|\sigma|^{p/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^\top(\mathbf{x}-\boldsymbol{\mu})\right), \tag{9}$$

and this is termed a *spherical Gaussian density*.

Remark. In (9),

$$(\mathbf{x}-\boldsymbol{\mu})^\top(\mathbf{x}-\boldsymbol{\mu}) = a^2$$

is the equation of a p-dimensional sphere centered at $\boldsymbol{\mu}$; in other words, the equation $(\mathbf{x}-\boldsymbol{\mu})^\top(\mathbf{x}-\boldsymbol{\mu}) = a^2$ is an *ellipsoid* centered at $\boldsymbol{\mu}$.

In general, the equation

$$(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = a^2$$

is an ellipsoid centered at $\boldsymbol{\mu}$, with $\boldsymbol{\Sigma}$ determining its orientation and shape. The multivariate Gaussian density function is *constant* along these ellipsoids.

7. **2-dimensional Gaussian Random Vector:** Let $p = 2$ and $X = (X_1, X_2)^\top \sim \text{Normal}_2(\mu, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\mu_1, \mu_2)^\top, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

σ_1^2 is the variance of X_1, σ_2^2 is the variance of X_2, and

$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}[X_1]\,\text{Var}[X_2]}} = \frac{\sigma_{1,2}}{\sigma_1\sigma_2}$$

is the correlation between X_1 and X_2. It follows that

$$|\boldsymbol{\Sigma}| = (1-\rho^2)\sigma_1^2\sigma_2^2,$$

and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \sigma_1^{-2} & -\rho\sigma_1^{-1}\sigma_2^{-1} \\ -\rho\sigma_1^{-1}\sigma_2^{-1} & \sigma_2^{-2} \end{pmatrix}.$$

The density function of the resulting bivariate Gaussian random vector is

$$f(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}Q\right),$$

where

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right].$$

If X_1 and X_2 are uncorrelated, $\rho = 0$, and the bivariate Gaussian density function reduces to the product of two univariate Gaussian densities,

$$\begin{aligned} f(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right) \\ &= f(x_1 \,|\, \mu_1, \sigma_1^2) \times f(x_2 \,|\, \mu_2, \sigma_2^2), \end{aligned}$$

implying that X_1 and X_2 are independent.

8. **"Partitioned" Gaussian Distribution:** Consider two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ and let $\mathbf{Z}$ be the $(p+q)$-dimensional random vector

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix} \in \mathbb{R}^{p+q}.$$

Assume that Z has a multivariate Gaussian distribution, and then, the exponent in the density function is the following quadratic form

$$-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_Z)^\top \boldsymbol{\Sigma}_Z^{-1}(\mathbf{z} - \boldsymbol{\mu}_Z).$$

The inverse matrix of $\boldsymbol{\Sigma}_Z$ is

$$\boldsymbol{\Sigma}_Z^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where

$$\begin{aligned}
\mathbf{A}_{11} &= \boldsymbol{\Sigma}_{XX}^{-1} + \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}(\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY})^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}, \\
\mathbf{A}_{12} &= -\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}(\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY})^{-1} = \mathbf{A}_{21}^\top, \\
\mathbf{A}_{22} &= (\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY})^{-1}.
\end{aligned}$$

In particular, we can write $\boldsymbol{\Sigma}_{ZZ}^{-1}$ as

$$\begin{pmatrix} \mathbf{I}_p & -\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{XX}^{-1} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ -\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1} & \mathbf{I}_q \end{pmatrix}.$$

9. **Transformation of Gaussian Random Vector:** Consider the following nonsingular transformation of $\mathbf{Z}$

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ -\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}.$$

Then, the mean of U is

$$\boldsymbol{\mu}_U = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ -\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1} & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix},$$

and the covariance matrix is

$$\boldsymbol{\Sigma}_{UU} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} \end{pmatrix}.$$

Therefore,

- the marginal distribution of $U_1 = X$ is $\text{Normal}_p(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_{XX})$,
- the marginal distribution of $U_2 = Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}X$ is

$$\text{Normal}_q(\mu_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\mu_X, \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}),$$

  and

- U_1 and U_2 are independent.

10. Conditional Gaussian Distribution: Given $X = \mathbf{x} \in \mathbb{R}^p$, the *conditional distribution* of Y is a q-variate Gaussian distribution with mean vector and covariance matrix given by

$$\boldsymbol{\mu}_{Y|\mathbf{x}} = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\mathbf{x} - \mu_X),$$
$$\boldsymbol{\Sigma}_{Y|\mathbf{x}} = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY},$$

respectively.

IV. Random Matrices

1. Random Matrix: The $m \times n$ matrix

$$\mathbf{Z} = \begin{pmatrix} Z_{1,1} & Z_{1,2} & \cdots & Z_{1,n} \\ Z_{2,1} & Z_{2,2} & \cdots & Z_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m,1} & Z_{m,2} & \cdots & Z_{m,n} \end{pmatrix}$$

with m rows and n columns is a *matrix-valued random variable* if each entry $Z_{i,j}$ is a random variable for all $i = 1, \cdots, m$ and $j = 1, \cdots, n$.

2. Expected Value of a Random Matrix:

$$\boldsymbol{\mu}_{\mathbf{Z}} = \mathbb{E}[\mathbf{Z}] = \begin{pmatrix} \mathbb{E}[Z_{1,1}] & \mathbb{E}[Z_{1,2}] & \cdots & \mathbb{E}[Z_{1,n}] \\ \mathbb{E}[Z_{2,1}] & \mathbb{E}[Z_{2,2}] & \cdots & \mathbb{E}[Z_{2,n}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[Z_{m,1}] & \mathbb{E}[Z_{m,2}] & \cdots & \mathbb{E}[Z_{m,n}]. \end{pmatrix}$$

3. Covariance Matrix of a Random Matrix: The *covariance matrix* of a random matrix $\mathbf{Z}$ is the matrix of covariances of all pairs of elements in $\mathbf{Z}$, i.e.,

$$\boldsymbol{\Sigma}_{\mathbf{ZZ}} = \mathrm{Cov}\big(\mathrm{vec}(\mathbf{Z}), \mathrm{vec}(\mathbf{Z})\big) = \mathbb{E}[\mathrm{vec}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})\,\mathrm{vec}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})^\top] \in \mathbb{R}^{mn \times mn}.$$

4. Transformation of $\mathbf{Z} \mapsto \mathbf{W} = \mathbf{A}\mathbf{Z}\mathbf{B}^\top + \mathbf{C}$: Consider the following transformation of

$$\mathbf{Z} \mapsto \mathbf{W} = \mathbf{A}\mathbf{Z}\mathbf{B}^\top + \mathbf{C},$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are constant matrices. Then,

(a) $\boldsymbol{\mu}_{\mathbf{W}} = \mathbb{E}[\mathbf{A}\mathbf{Z}\mathbf{B}^\top + \mathbf{C}] = \mathbf{A}\,\mathbb{E}[\mathbf{Z}]\mathbf{B}^\top + \mathbf{C} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{Z}}\mathbf{B}^\top + \mathbf{C}$;

(b) $\boldsymbol{\Sigma}_{\mathbf{WW}} = \mathrm{Var}[\mathbf{A}\mathbf{Z}\mathbf{B}^\top + \mathbf{C}] = \mathrm{Var}[\mathbf{A}\mathbf{Z}\mathbf{B}^\top] = \mathbb{E}[\mathrm{vec}(\mathbf{W} - \boldsymbol{\mu}_{\mathbf{W}})\,\mathrm{vec}(\mathbf{W} - \boldsymbol{\mu}_{\mathbf{W}})^\top]$.
Derivation: Since

$$\mathrm{vec}(\mathbf{W} - \boldsymbol{\mu}_{\mathbf{W}}) = \mathrm{vec}(\mathbf{A}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})\mathbf{B}^\top) = (\mathbf{A} \otimes \mathbf{B})\,\mathrm{vec}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}}),$$

it follows that

$$\begin{aligned}\mathbf{\Sigma_{WW}} &= \mathbb{E}[(\mathbf{A}\otimes\mathbf{B})\operatorname{vec}(\mathbf{Z}-\boldsymbol{\mu}_\mathbf{Z})((\mathbf{A}\otimes\mathbf{B})\operatorname{vec}(\mathbf{Z}-\boldsymbol{\mu}_\mathbf{Z}))^\top] \\ &= (\mathbf{A}\otimes\mathbf{B})\,\mathbb{E}[\operatorname{vec}(\mathbf{Z}-\boldsymbol{\mu}_\mathbf{Z})\operatorname{vec}(\mathbf{Z}-\boldsymbol{\mu}_\mathbf{Z})^\top](\mathbf{A}\otimes\mathbf{B})^\top \\ &= (\mathbf{A}\otimes\mathbf{B})\mathbf{\Sigma_{ZZ}}(\mathbf{A}^\top\otimes\mathbf{B}^\top).\end{aligned}$$

5. Wishart Distribution:

(a) *Definition:* Let X_i, $i = 1, \cdots, n$, be n independent p-dimensional random vectors distributed as

$$X_i \sim \text{Normal}_p(\boldsymbol{\mu}_i, \mathbf{\Sigma}), \qquad \text{for all } i = 1, \cdots, n \geq p.$$

Define the following $p \times p$ positive semidefinite random matrix

$$\mathbf{W} := \sum_{i=1}^n X_i X_i^\top.$$

Then, $\mathbf{W}$ is said to have the *Wishart distribution* with n degrees of freedom and associated matrix $\mathbf{\Sigma}$, denoted by $\mathbf{W} \sim \text{Wishart}_p(n, \mathbf{\Sigma})$.

If $\boldsymbol{\mu}_i = \mathbf{0}_p$, the resulting Wishart random matrix $\mathbf{W}$ is said to be *central*; otherwise, it is said to be *non-central*.

(b) *Density Function:* The joint density function of the $p(p+1)/2$ elements of $\mathbf{W}$ is

$$f_\mathbf{W}(\mathbf{w}\,|\,n, \mathbf{\Sigma}) = c_{p,n}|\mathbf{\Sigma}|^{-n/2}|\mathbf{w}|^{\frac{1}{2}(n-p-1)}\exp\left(-\frac{1}{2}\operatorname{trace}(\mathbf{w}\mathbf{\Sigma}^{-1})\right),$$

where

$$\frac{1}{c_{p,n}} = 2^{\frac{np}{2}}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^p \Gamma\left(\frac{n+1-i}{2}\right).$$

Remark 1. If $\mathbf{W}$ is singular, the density is 0, and the corresponding Wishart random matrix $\mathbf{W}$ is said to be *singular*.

Remark 2. If $p = 1$, $\text{Wishart}_1(n, \sigma^2)$ is identical to a $\sigma^2\chi_n^2$ distribution.

(c) *Moments:* The first two moments of the Wishart distribution $\text{Wishart}_p(n, \mathbf{\Sigma})$ are

$$\begin{aligned}\mathbb{E}[\mathbf{W}] &= n\mathbf{\Sigma}, \\ \operatorname{Var}[\operatorname{vec}(\mathbf{W})] &= \mathbb{E}\Big[\big(\operatorname{vec}(\mathbf{W}-n\mathbf{\Sigma})\big)\operatorname{vec}(\mathbf{W}-n\mathbf{\Sigma})^\top\Big] \\ &= n(\mathbf{I}_{p^2} + \mathbf{I}_{(p,p)})(\mathbf{\Sigma}\otimes\mathbf{\Sigma}),\end{aligned}$$

where $\mathbf{I}_{(p,q)}$ is a *permuted identity matrix* and is a $pq \times pq$-matrix partitioned into $(p\times q)$-submatrices such that the (i,j)-th submatrix has a 1 in its (j,i)-th position and zeros everywhere else.

(d) *Properties of Wishart Distribution:*

i. Let $\mathbf{W}_j \sim \text{Wishart}_p(n_j, \mathbf{\Sigma})$, $j = 1, 2, \cdots, m$, be independently distributed (central or not). Then,

$$\sum_{j=1}^{n} \mathbf{W}_j \sim \text{Wishart}_p\left(\sum_{j=1}^{m} n_j, \mathbf{\Sigma}\right).$$

ii. Suppose $\mathbf{W} \sim \text{Wishart}_p(n, \mathbf{\Sigma})$, and let $\mathbf{A} \in \mathbb{R}^{d\times p}$ be a constant matrix with rank d. Then,

$$\mathbf{A}\mathbf{W}\mathbf{A}^\top \sim \text{Wishart}_d(n, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top).$$

iii. Suppose $\mathbf{W} \sim \text{Wishart}_p(n, \mathbf{\Sigma})$, and let $\mathbf{v} \in \mathbb{R}^p$ be a fixed vector. Then,

$$\mathbf{v}^\top \mathbf{W}\mathbf{v} \sim \sigma_{\mathbf{v}}^2 \chi_n^2,$$

where $\sigma_{\mathbf{v}}^2 := \mathbf{v}^\top \mathbf{\Sigma}\mathbf{v}$. In particular, the chi-squared distribution is central if the Wishart distribution is central.

iv. Let $\mathbf{X} = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n\times p}$, where $X_i \sim \text{Normal}_p(\mathbf{0}_p, \mathbf{\Sigma})$, for $i = 1, 2, \cdots, n$, are independently and identically distributed. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric with rank r, and let $\mathbf{v} \in \mathbb{R}^p$ be a fixed vector. Let $\mathbf{y} = \mathbf{X}\mathbf{v}$. Then,

$$\mathbf{X}^\top \mathbf{A}\mathbf{X} \sim \text{Wishart}_p(r, \mathbf{\Sigma})$$

if and only if $\mathbf{y}^\top \mathbf{A}\mathbf{y} \sim \sigma_{\mathbf{v}}^2 \chi_n^2$, where $\sigma_{\mathbf{v}}^2 := \mathbf{v}^\top \mathbf{\Sigma}\mathbf{v}$.

6. Properties of Permuted Identity Matrix:

(a) The permuted identity matrix $I_{(p,p)}$ can be expressed as the sum of p^2 Kronecker products as

$$I_{(p,p)} = \sum_{i=1}^{p}\sum_{j=1}^{p} (\mathbf{H}_{i,j} \otimes \mathbf{H}_{i,j}^\top),$$

where $\mathbf{H}_{i,j} \in \mathbb{R}^{p\times p}$ is a matrix with (i, j)-th element equal to 1 and zero otherwise.

(b) For any $\mathbf{A} \in \mathbb{R}^{p\times p}$, we have

$$I_{(p,p)}\, \text{vec}(\mathbf{A}) = \text{vec}(\mathbf{A}^\top).$$

V. Maximum Likelihood Estimation of the Gaussian Random Vector

1. Setup: Assume that $X_1, X_2, \cdots, X_n$ are n i.i.d p-dimensional Gaussian random vectors, that is,

$$X_i \overset{\text{i.i.d}}{\sim} \text{Normal}_p(\boldsymbol{\mu}, \mathbf{\Sigma}), \qquad \text{for all } i = 1, \cdots, n,$$

where the parameters, $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, are both *unknown*. We estimate $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ using the method of maximum likelihood.

2. **Likelihood Function:** By independence, the *likelihood function* of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is

$$L\big(\boldsymbol{\mu}, \boldsymbol{\Sigma} \,|\, X_1, \cdots, X_n\big) = (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(X_i - \boldsymbol{\mu})\right),$$

and the *log-likelihood function* is

$$\begin{aligned}
\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &:= \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \,|\, X_1, \cdots, X_n) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(X_i - \boldsymbol{\mu}) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\operatorname{trace}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^\top (X_i - \bar{X})\right) \\
&\qquad - \frac{n}{2}(\bar{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{X} - \boldsymbol{\mu}),
\end{aligned}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the *sample mean*.

3. **MLE of $\boldsymbol{\mu}$:** To find the MLE of $\boldsymbol{\mu}$, we differentiate $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\mu}$ and obtain

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = n\boldsymbol{\Sigma}^{-1}(\bar{X} - \boldsymbol{\mu}).$$

Setting this derivative to 0, the MLE of $\boldsymbol{\mu}$ is

$$\widehat{\boldsymbol{\mu}} = \bar{X},$$

the *sample mean*.

4. **MLE of $\boldsymbol{\Sigma}$:** Plugging $\widehat{\boldsymbol{\mu}} = \bar{X}$ back into $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have

$$\ell(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\operatorname{trace}\Big(\boldsymbol{\Sigma}^{-1}\mathbf{S}\Big),$$

where $\mathbf{S} := \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top$.

We take the derivative of $\ell(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ and obtain

$$\frac{\partial \ell}{\partial \boldsymbol{\Sigma}}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = -\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}.$$

Setting this derivative to $\mathbf{0}_{p\times p}$, we have the MLE of $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{S},$$

the *sample covariance matrix*.

5. **Unbiased of $\widehat{\mu}$ and $\widehat{\boldsymbol{\Sigma}}$:**

(a) The MLE of $\boldsymbol{\mu}$, $\widehat{\boldsymbol{\mu}} = \bar{X}$, is an unbiased estimator of μ, that is,

$$\mathbb{E}[\bar{X}] = \boldsymbol{\mu};$$

(b) The MLE of $\boldsymbol{\Sigma}$, $\widehat{\boldsymbol{\Sigma}} = (1/n)\mathbf{S}$, is *not* unbiased, and

$$\mathbb{E}\big[\widehat{\boldsymbol{\Sigma}}\big] = \frac{n-1}{n}\boldsymbol{\Sigma}.$$

6. Sampling Distribution of $\widehat{\boldsymbol{\mu}} = \bar{X}$: Since $\bar{X}$ is a linear combination of $X_1, \cdots, X_n$, each of which is i.i.d as $\text{Normal}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\widehat{\boldsymbol{\mu}} = \bar{X}$ is distributed as

$$\bar{X} \sim \text{Normal}_p\bigg(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\bigg).$$

7. Sampling Distribution of $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{S}$:

(a) *Assuming $\boldsymbol{\mu} = \mathbf{0}_p$:* Let $\mathbf{v} \in \mathbb{R}^p$ be a fixed vector and consider $Y_i = \mathbf{v}^\top X_i$, for all $i = 1, 2, \cdots, n$. Then,

$$Y_i \sim \text{Normal}_1(0, \sigma_{\mathbf{v}}^2), \qquad \text{where } \sigma_{\mathbf{v}}^2 = \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v},$$

and

$$Y := (Y_1, Y_2, \cdots, Y_n)^\top \sim \text{Normal}_n(\mathbf{0}_n, \sigma_{\mathbf{v}}^2 \cdot \mathbf{I}_n).$$

Let $\mathbf{A} = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$, where $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^\top$ is a matrix with all entries being 1. Note that $\mathbf{A}$ is idempotent with rank $n-1$. From univariate theory,

$$\frac{1}{n}\mathbf{1}_n^\top Y = \bar{Y} \sim \text{Normal}_1\bigg(0, \frac{1}{n}\sigma_{\mathbf{v}}^2\bigg),$$

and

$$Y^\top \mathbf{A} Y = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma_{\mathbf{v}}^2 \cdot \chi_{n-1}^2$$

are independently distributed for any $\mathbf{v}$.

Now, let $\mathbf{X} = (X_1, \cdots, X_n)^\top$. Then,

$$\frac{1}{n}\mathbf{X}^\top \mathbf{1}_n \sim \text{Normal}_p\bigg(\mathbf{0}_p, \frac{1}{n}\boldsymbol{\Sigma}\bigg),$$

and, using the properties of the Wishart distribution,

$$\mathbf{X}^\top \mathbf{A} \mathbf{X} = \mathbf{S} \sim \text{Wishart}_p(n-1, \boldsymbol{\Sigma}). \tag{10}$$

Independence of $\bar{X}$ and $\mathbf{S}$*:* Because $Y \sim \text{Normal}_n(\mathbf{0}_n, \sigma_{\mathbf{v}}^2 \cdot \mathbf{I}_n)$, it follows that

$$\frac{1}{n}\mathbf{1}_n^\top Y \sim \text{Normal}_1\left(0, \frac{1}{n}\sigma_{\mathbf{v}}^2\right), \qquad \text{and} \qquad Y^\top \mathbf{J}_n Y \sim \sigma_{\mathbf{v}}^2 \cdot \chi_n^2.$$

Furthermore, it is easy to obtain $\mathbf{A}(\frac{1}{n}\mathbf{1}_n) = \mathbf{0}_n$ so that the columns of $\mathbf{A}$ and $\frac{1}{n}\mathbf{1}_n$ are mutually orthogonal. Thus,

$$\mathbf{X}^\top \mathbf{a}_i = X_i - \bar{X}, \qquad \text{for all } i = 1, 2, \cdots, n,$$

where $\mathbf{a}_i$ is the i-th column of $\mathbf{A}$, and $\mathbf{X}^\top(\frac{1}{n}\mathbf{1}_n)$ are statistically independent of each other. Thus,

$$\mathbf{X}^\top\left(\frac{1}{n}\mathbf{1}_n\right) = \bar{X} \qquad \text{and} \qquad \mathbf{X}^\top \mathbf{A} \mathbf{X} = (\mathbf{X}^\top \mathbf{A})(\mathbf{X}^\top \mathbf{A})^\top = \mathbf{S}$$

are independently distributed.

(b) *Assuming* $\boldsymbol{\mu} \neq \mathbf{0}_p$*:* The case of $\boldsymbol{\mu} \neq \mathbf{0}_p$ is dealt with by replacing X_i by $X_i - \mu$, for $i = 1, 2, \cdots, n$. This does *not* change $\mathbf{S}$, and $\bar{X}$ above is replaced by $\bar{X} - \boldsymbol{\mu}$. Thus, $\mathbf{S}$ is independent of $\bar{X} - \boldsymbol{\mu}$ (and, hence, of $\bar{X}$), and

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{S} \sim \frac{1}{n}\text{Wishart}_p(n-1, \boldsymbol{\Sigma}). \tag{11}$$

References

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.