Liangji's Notes for Linear Algebra

Liangji Li

October 8, 2024

Abstract

TODO: Here is where I would say something.

Contents

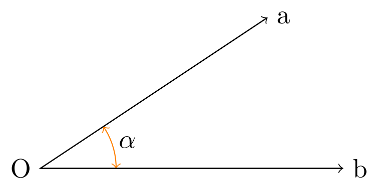


Figure 1: Inner product: $\mathbf{a}^\top \mathbf{b}$

1 Products of Two vectors

1.1 Inner Product

An inner product of **a** and **b** can be expressed in several ways:

1. $\langle \mathbf{a}, \mathbf{b} \rangle$
2. $\mathbf{a} \cdot \mathbf{b}$
3. $\mathbf{a}^\top \mathbf{b}$

Definition 1.1. L_2 **Norm:**

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{\|\mathbf{x}\|^2} \tag{1.1}$$

Actually, the L_2 norm can also be considered as Euclidean distance (length).

An inner product can be expressed in terms of lengths and the angle between them.

$$\mathbf{a}^\top \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \alpha \tag{1.2}$$

where α is the angle between **a** and **b**, as shown in figure 1. Specially, if $\|\mathbf{a}\| = 1$, $\mathbf{a}^\top \mathbf{b}$ is said **the coordinate of b relative to a**.

Theorem 1.1. Cauchy-Schwarz Inequality Give two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then

$$(\mathbf{x}^\top \mathbf{y})^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \tag{1.3}$$

Proof. Let c $= \dfrac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}}$. If $\mathbf{x}$ is $\mathbf{0}$, then the proof is immediately completed, and therefore suppose $\mathbf{x} \neq \mathbf{0}$. Expanding

$$\begin{aligned}
\|\mathbf{y} - c\mathbf{x}\|^2 &= \mathbf{y}^\top \mathbf{y} - 2c\mathbf{x}^\top \mathbf{y} + c^2 \mathbf{x}^\top \mathbf{x} && (1.4)\\
&= \mathbf{y}^\top \mathbf{y} - 2\frac{(\mathbf{x}^\top \mathbf{y})^2}{\mathbf{x}^\top \mathbf{x}} + \frac{(\mathbf{x}^\top \mathbf{y})^2}{\mathbf{x}^\top \mathbf{x}} && (1.5)\\
&= \|\mathbf{y}\|^2 - \frac{(\mathbf{x}^\top \mathbf{y})^2}{\|\mathbf{x}\|^2} \geq 0 && (1.6)
\end{aligned}$$

□

Theorem 1.2. If $(\mathbf{x}^\top \mathbf{y})^2 = \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$, then $\mathbf{x}, \mathbf{y}$ are linearly dependent.

Proof. By the equation (1.6), if $(\mathbf{x}^\top \mathbf{y})^2 = \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$, then $\|\mathbf{y} - c\mathbf{x}\|^2 = 0$, implying $\mathbf{y} - c\mathbf{x} = \mathbf{0}$. Thus, $\mathbf{y} = c\mathbf{x}$. □

1.2 Outer Product (Tensor Product)

An outer product takes as inputs two vectors and then produces a matrix:

$$\mathbf{a}\mathbf{b}^\top = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \mathbf{b}^\top = \begin{bmatrix} a_1\mathbf{b}^\top \\ \vdots \\ a_n\mathbf{b}^\top \end{bmatrix}$$

It can also be denoted by $\mathbf{a} \otimes \mathbf{b}$.

2 Views of Matrix Multiplication

2.1 Linear Combination of Columns

Given two matrices, $A_{n\times p}$ and $B_{p\times m}$, each column of their product can be expressed as a linear combination of the columns of A.

$$AB = \begin{bmatrix} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_m \end{bmatrix} \tag{2.1}$$

2.2 Sum of outer products

AB can be expressed as a sum of outer products of $\mathbf{a}_i\mathbf{b}^{(i)}$, where $\mathbf{b}^{(i)}$ is the i^{th} of B.

$$AB = \sum_{i=1}^{p} \mathbf{a}_i\mathbf{b}^{(i)} \tag{2.2}$$

Notice that $\mathbf{a}_i\mathbf{b}^{(i)}$ is of rank 1 matrix, since each column of $\mathbf{a}_i\mathbf{b}^{(i)}$ is a multiple of $\mathbf{a}_i$.

2.3 TODO: Linear Combination of Rows

3 Gram Matrix

3.1 Information carried by Gram Matrix

In the most cases, the data matrix, $A \in \mathbb{R}^{n\times p}$, is not square, and thus its inverse does not exist. For convenience of computation, we can "reduce" the data matrix into a square matrix:

$$A^\top A = \begin{bmatrix} \mathbf{a_1}^\top\mathbf{a_1} & \mathbf{a_1}^\top\mathbf{a_2} & \cdots & \mathbf{a_1}^\top\mathbf{a_p} \\ \vdots & \vdots & \ldots & \vdots \\ \mathbf{a_p}^\top\mathbf{a_1} & \mathbf{a_p}^\top\mathbf{a_2} & \cdots & \mathbf{a_p}^\top\mathbf{a_p} \end{bmatrix} = \begin{bmatrix} \|\mathbf{a}_1\|\|\mathbf{a}_1\|\cos\theta_{1,1} & \|\mathbf{a}_1\|\|\mathbf{a}_2\|\cos\theta_{1,2} & \cdots & \|\mathbf{a}_1\|\|\mathbf{a}_p\|\cos\theta_{1,p} \\ \vdots & \vdots & \ldots & \vdots \\ \|\mathbf{a}_p\|\|\mathbf{a}_1\|\cos\theta_{p,1} & \|\mathbf{a}_p\|\|\mathbf{a}_2\|\cos\theta_{p,2} & \cdot & \|\mathbf{a}_p\|\|\mathbf{a}_p\|\cos\theta_{p,p} \end{bmatrix} \tag{3.1}$$

Suppose that n is larger than p, we can reduce A into a relatively small matrix $G \in \mathbb{R}^{p\times p}$ which contains the necessary information about the columns vector of A. The necessary information of a column vector of A consists of its length and its angles with the other column vectors, which is contained by G.

Remark 3.1. Although G contains the necessary information for the column vectors of X, we cannot use this information to directly restore X from G. However, we can find a collection of vectors with the same relations as those between the column vectors of A, by using a matrix called **cosine similarity matrix** and the **Choleskey Decomposition**.

3.2 Cosine Similarity Matrix

If we let S be a diagonal matrix

$$S = \begin{bmatrix} \|\mathbf{a}_1\| & 0 & \cdots & 0 \\ 0 & \|\mathbf{a}_2\| & \cdots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \cdots & \|\mathbf{a}_p\| \end{bmatrix}$$

We can get the **Cosine Similarity Matrix** C for G:

$$C = S^{-1}GS = \begin{bmatrix} 1 & \cos\theta_{1,2} & \cdots & \cos\theta_{1,p} \\ \vdots & \vdots & \ldots & \vdots \\ \cos\theta_{p,1} & \cdots & \cos\theta_{p,2} & \cdot & 1 \end{bmatrix}$$

If the angles $\theta_{i,j}$ satisfies some conditions (TODO), C can be Cholesky-decomposed into

$$C = R^\top R$$

where the columns of R are **unit vectors that can reflect the relations between the column vectors of** X. C and G are called **similar** to each other by the definition 9.3.

4 Coordinate Systems

4.1 Coordinates relative to a Basis

Theorem 4.1. The Unique Representation Theorem: Let $\mathcal{B} = \{\mathbf{b}_1, \cdots, \mathbf{b}_n\}$ be a basis for a vector space V. Then $\forall \mathbf{x} \in V$, there exists a unique set of scalars $c_1, \cdots, c_n$ such that

$$\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n$$

Definition 4.1. $\mathcal{B}$-coordinates: Suppose $\mathcal{B} = \{\mathbf{b}_1, \cdots, \mathbf{b}_n\}$ is a basis for V and $\mathbf{x} \in V$. The **coordinates of x relative to the basis $\mathcal{B}$ (or shortly coordinates of $\mathcal{B}$** are the weights $c_1, \cdots, c_n$ such that

$$\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n$$

It is denoted by

$$[\mathbf{x}]_\mathcal{B} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$$

Remark 4.1. It is easy to see that $[\cdot]_\mathcal{B}$ is a linear transformation, that is:

$$[c\mathbf{a} + \mathbf{b}]_\mathcal{B} = c[\mathbf{a}]_\mathcal{B} + [\mathbf{b}]_\mathcal{B}$$

In fact, for any vector $\mathbf{x}$ in $\mathbb{R}^n$, its $\mathcal{E}$-coordinate is itself, where $\mathcal{E}$ is standard basis

$$[\mathbf{x}]_\mathcal{E} = \mathbf{x}$$

4.2 Change of Coordiantes

Definition 4.2. Change-of-Coordinates Matrix: Let

$$P_\mathcal{B} = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{bmatrix}$$

Then the vector equation $\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n$ is equivalent to

$$\mathbf{x} = P_\mathcal{B}[\mathbf{x}]_\mathcal{B}$$

$P_\mathcal{B}$ is called **change-of-coordinates matrix** from **B** to **the standard basis** $\mathcal{E}$ in $\mathbb{R}^n$. Since $\mathcal{B}$ is a basis in $\mathbb{R}^n$, its inverse $P_\mathcal{B}^{-1}$ always exists. Left-multiplication by $P_\mathcal{B}^{-1}$ converts $\mathbf{x}$ into its $\mathcal{B}$-coordinate vector

$$P_\mathcal{B}^{-1}\mathbf{x} = [\mathbf{x}]_\mathcal{B}$$

Theorem 4.2. Let $\mathcal{B} = \{\mathbf{b}_1, \cdots, \mathbf{b}_n\}$ and $\mathcal{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_n\}$ be bases of a vector space V. Then there is a unique $n \times n$ matrix $\underset{\mathcal{C} \leftarrow \mathcal{B}}{P}$ such that

$$[\mathbf{x}]_{\mathcal{C}} = \underset{\mathcal{C} \leftarrow \mathcal{B}}{P} [\mathbf{x}]_{\mathcal{B}}$$

The columns of $\underset{\mathcal{C} \leftarrow \mathcal{B}}{P}$ are the $\mathcal{C}$-coordinate vectors of the vectors in the basis $\mathcal{B}$. That is,

$$\underset{\mathcal{C} \leftarrow \mathcal{B}}{P} = \begin{bmatrix} [\mathbf{b}_1]_{\mathcal{C}} & [\mathbf{b}_2]_{\mathcal{C}} & \cdots & [\mathbf{b}_n]_{\mathcal{C}} \end{bmatrix}$$

The matrix $\underset{\mathcal{C} \leftarrow \mathcal{B}}{P}$ is called **change-of-coordinates matrix from $\mathcal{B}$ to $\mathcal{C}$**. That is,

$$[\mathbf{x}]_{\mathcal{C}} = \underset{\mathcal{C} \leftarrow \mathcal{B}}{P} [\mathbf{x}]_{\mathcal{B}}$$

Similarly, the inverse of $\underset{\mathcal{C} \leftarrow \mathcal{B}}{P}$ always exists

$$(\underset{\mathcal{C} \leftarrow \mathcal{B}}{P})^{-1} = \underset{\mathcal{B} \leftarrow \mathcal{C}}{P}$$

Note that $P_{\mathcal{B}}$ implies that $\underset{\mathcal{E} \leftarrow \mathcal{B}}{P}$. One of the ways to calculate $\underset{\mathcal{C} \leftarrow \mathcal{B}}{P}$ is to place the two sets of bases into a matrix, and then solve it as if it were a simple linear equation:

$$\begin{bmatrix} \mathcal{C} \mid \mathcal{B} \end{bmatrix} \sim [\ I \mid \underset{\mathcal{C} \leftarrow \mathcal{B}}{P}\ ]$$

5 Orthogonality

Definition 5.1. Two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are **orthogonal** to each other if $\mathbf{u}^\top \mathbf{v} = 0$ or $\mathbf{v}^\top \mathbf{u} = 0$,

5.1 Orthogonal Complement

If a vector $\mathbf{v}$ is orthogonal to every vector in a subspace W of $\mathbb{R}^n$, then $\mathbf{v}$ is said to be **orthogonal to** W. The set of all vectors $\mathbf{v}$ that are orthogonal to W is called the **orthogonal complement** of W.

Definition 5.2. Orthogonal Complement: A subspace V is the orthogonal complement of W, if

$$W^\perp = \{\mathbf{v} \in V \mid \forall \mathbf{u} \in W : \mathbf{v}^\top \mathbf{u}\}$$

Definition 5.3. Direct Sum: Let W_1 and W_2 be subspaces of a vector space V, if

$$\forall \mathbf{v} \in V : \mathbf{v} = \underbrace{\mathbf{w}_1 + \mathbf{w}_2}_{\text{uniquely}} \quad \text{where } \mathbf{w}_1 \in W_1, \mathbf{w}_2 \in W_2$$

then V is called the **direct sum** of W_1 and W_2. In this case, we write $V = W_1 \oplus W_2$.

Theorem 5.1. If $V = W_1 \oplus W_2$, then $W_1 \cap W_2 = \{\mathbf{0}\}$.

Proof. Let $\mathbf{v} \in W_1 \cap W_2$. Since $\mathbf{v}$ is also in V. Then

$$\mathbf{v} = \mathbf{0} + \mathbf{w_1} \quad \text{and} \quad \mathbf{v} = \mathbf{0} + \mathbf{w}_2$$

with $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$. By the uniqueness of direct sum representations, we have $\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{0}$. □

Theorem 5.2. If W is a subspace of an inner product space V, then

$$V = W \oplus W^\perp \quad \text{and} \quad \dim(V) = \dim(W_1) + \dim(W_2).$$

Theorem 5.3. Let A be an $m \times n$ matrix, then

$$\Big(\text{Row}(A)\Big)^{\perp} = \text{Nul}(A) \quad \text{and} \quad \Big(\text{Col}(A)\Big)^{\perp} = \text{Nul}(A^{\top}) \tag{5.1}$$

By theorem 5.2, it is clear that

$$\dim\Big(\text{Row}(A)\Big) + \dim\Big(\text{Nul}(A)\Big) = m \quad \text{and} \quad \dim\Big(\text{Col}(A)\Big) + \dim\Big(\text{Nul}(A^{\top})\Big) = n \tag{5.2}$$

5.2 Orthogonal Projection

The orthogonal projection of $\mathbf{y}$ on $\mathbf{x}$ can be expressed as

$$\text{proj}_{\mathbf{x}}(\mathbf{y}) = \frac{\mathbf{x}^{\top}\mathbf{y}}{\mathbf{x}^{\top}\mathbf{x}}\mathbf{x} \tag{5.3}$$

The equation (5.3) can be written in a matrix-vector multiplication form:

$$\text{proj}_{\mathbf{x}}(\mathbf{y}) = \frac{\mathbf{x}(\mathbf{x}^{\top}\mathbf{y})}{\|\mathbf{x}\|^2} = \frac{(\mathbf{x}\mathbf{x}^{\top})\mathbf{y}}{\|\mathbf{x}\|^2} = \left(\frac{\mathbf{x}}{\|\mathbf{x}\|} \otimes \frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\mathbf{y} \tag{5.4}$$

$\dfrac{\mathbf{x}}{\|\mathbf{x}\|} \otimes \dfrac{\mathbf{x}}{\|\mathbf{x}\|}$ is called **Projection Matrix**.

Example 5.1. Given two vectors $\mathbb{1}, \mathbf{y} \in \mathbb{R}^n$, calculate the projection of $\mathbf{y}$ onto $\mathbb{1}$.

Solution. Calculate the projection matrix

$$\frac{\mathbb{1}}{\|\mathbb{1}\|} \otimes \frac{\mathbb{1}}{\|\mathbb{1}\|} = \frac{\mathbb{1} \otimes \mathbb{1}}{n}$$

The projection vector of $\mathbf{y}$ onto $\mathbb{1}$ is given by

$$\frac{\mathbb{1} \otimes \mathbb{1}}{n}\mathbf{y} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \vdots \\ \sum_{i=1}^{n} y_i \end{bmatrix} = \overline{y}\ \mathbb{1}$$

That is, the projection vector of $\mathbf{y}$ onto $\mathbb{1}$ is called **sample mean vector of y**.

Remark 5.1. The project matrix of $\mathbb{1}$ is, in statistics, typically denoted by

$$H_0 = \mathbb{1}(\mathbb{1}^{\top}\mathbb{1})^{-1}\mathbb{1}^{\top} \tag{5.5}$$

The **Total Sum of Squares** in a linear model is defined as:

$$\text{SST} = \|\mathbf{y} - H_0\mathbf{y}\|^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{5.6}$$

Example 5.2. Let $X = (\mathbf{x}_1 \ \cdots \ \mathbf{x}_n)^\top$, we can calculate the projection scalar of $\mathbf{x}_i^\top$ onto a unit vector $\mathbf{v}$

$$\alpha = X\mathbf{v} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{v} \\ \vdots \\ \mathbf{x}_n^\top \mathbf{v} \end{bmatrix} \tag{5.7}$$

And we then can calculate the projection vectors on the unit vector $\mathbf{v}$

$$Z = X\mathbf{v}\mathbf{v}^\top = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{v}\mathbf{v}^\top \\ \vdots \\ \mathbf{x}_n^\top \mathbf{v}\mathbf{v}^\top \end{bmatrix} = XV = X(\mathbf{v} \otimes \mathbf{v}) \tag{5.8}$$

where V is the projection matrix of $\mathbf{v}$. Note that the i^{th} row, instead of the i^{th} column, of Z is the projection of $\mathbf{x}_i^\top$ on the unit vector $\mathbf{v}$.

5.3 Orthogonal Matrix

An orthogonal matrix V is **one that has an orthonormal set of vectors** as its columns. V has the following properties:

1. $V^\top V = I = VV^\top$
2. $V^\top = V^{-1}$
3. $V^\top$ is also an orthogonal matrix.
4. $\|V\mathbf{x}\|^2 = \|\mathbf{x}\|^2$

$VV^\top$ can be viewed as

$$VV^\top = \mathbf{v}_1 \otimes \mathbf{v}_1 + \cdots + \mathbf{v}_n \otimes \mathbf{v}_n = I \tag{5.9}$$

Note also that left-multiply $VV^\top$ by X

$$\begin{aligned} XVV^\top &= X(\mathbf{v}_1 \otimes \mathbf{v}_1 + \cdots + \mathbf{v}_n \otimes \mathbf{v}_n) && (5.10) \\ &= X\mathbf{v}_1 \otimes \mathbf{v}_1 + \cdots + X\mathbf{v}_n \otimes \mathbf{v}_n && (5.11) \\ &= XI = X && (5.12) \end{aligned}$$

Property 4 can be easily proved

Proof.

$$\|V\mathbf{x}\|^2 = (V\mathbf{x})^\top V\mathbf{x} = \mathbf{x}^\top V^\top V\mathbf{x} = \mathbf{x}^\top I\mathbf{x} = \|\mathbf{x}\|^2 \tag{5.13}$$

□

This property implies that **a linear transformation, whose transformation matrix is an orthogonal matrix, say $V^\top$, preserves the length and the angle**.

Theorem 5.4. If $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p\}$ is an orthonormal basis for a subspace W of $\mathbb{R}^n$, then

$$\text{proj}_W(\mathbf{y}) = (\mathbf{y}^\top \mathbf{u_1})\mathbf{u}_1 + (\mathbf{y}^\top \mathbf{u_2})\mathbf{u}_2 + \cdots + (\mathbf{y}^\top \mathbf{u_p})\mathbf{u}_p$$

Let $U = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_p \end{bmatrix}$ then

$$\forall \mathbf{y} \in \mathbb{R}^n : \text{proj}_W(\mathbf{y}) = UU^\top \mathbf{y} \tag{5.14}$$

5.4 The Gram-Schmidt Process and QR Factorization

The Gram-Schmidt process is a simple algorithm for producing an orthogonal or orthonormal basis for any nonzero subspace of $\mathbb{R}^n$.

Theorem 5.5. Given a basis $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p\}$ for a nonzero subspace W of $\mathbb{R}^n$, define

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{x}_1 \\
\mathbf{v}_2 &= \mathbf{x}_2 - \operatorname{proj}_{\mathbf{v}_1}(\mathbf{x}_2) \\
\mathbf{v}_3 &= \mathbf{x}_3 - \operatorname{proj}_{\mathbf{v}_1}(\mathbf{v}_3) - \operatorname{proj}_{\mathbf{v}_2}(\mathbf{x}_3) \\
&\vdots \\
\mathbf{v}_p &= \mathbf{x}_p - \operatorname{proj}_{\mathbf{v}_1}(\mathbf{x}_p) - \operatorname{proj}_{\mathbf{v}_2}(\mathbf{x}_p) - \cdots - \operatorname{proj}_{\mathbf{v}_{p-1}}(\mathbf{x}_p)
\end{aligned}$$

Then $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p\}$ is an orthogonal basis for W. In addition,

$$\operatorname{Span}\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p\} = \operatorname{Span}\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p\}$$

Remark 5.2. The theorem shows that any nonzero subspace W of $\mathbb{R}^n$ has an orthogonal basis. We can reduce the orthogonal basis into an orthonormal basis, $\mathcal{U} = \{\mathbf{v}'_1, \mathbf{v}'_2, \cdots, \mathbf{v}'_n\}$, by letting

$$\mathbf{v}'_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|}$$

Theorem 5.6. If A is an $m \times n$ matrix with linearly independent columns, then A can be factored as $A = QR$, where $Q \in \mathbb{R}^{m\times n}$ is a matrix whose columns form an **orthonormal basis** for $\operatorname{Col}(A)$ and $R \in \mathbb{R}^{n\times n}$ is an upper triangular non-singular matrix with positive entries on its diagonal.

Proof. Let $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ be a basis for $\operatorname{Col}(A)$. We can find a set of orthonormal basis $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ using Gram-Schmidt process. Let $Q = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n\end{bmatrix}$. Since $\mathbf{x}_k$ is in $\operatorname{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_k\} = \operatorname{Span}\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$, there exists $r_{1k}, \cdots, r_{kk}$ such that

$$\mathbf{x}_k = r_{1k}\mathbf{u}_1 + \cdots + r_{kk}\mathbf{u}_k + 0 \cdot \mathbf{u}_{k+1} + \cdots + 0 \cdot \mathbf{u}_n \tag{5.15}$$

We may assume that $r_{kk} > 0$. (If $r_{kk} < 0$, multiply both r_{kk} and $\mathbf{u}_k$ by -1.) Let

$$\mathbf{r}_k = \begin{bmatrix} r_{1k} & \cdots & r_{kk} & 0 & \cdots & 0\end{bmatrix}^\top$$

That is, $\mathbf{x}_k = Q\mathbf{r}_k$. Let $R = \begin{bmatrix}\mathbf{r}_1 & \cdots & \mathbf{r}_n\end{bmatrix}$. Then

$$A = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_n\end{bmatrix} = \begin{bmatrix}Q\mathbf{r}_1 & \cdots & Q\mathbf{r}_n\end{bmatrix} = QR$$

The fact that R is non-singular follows easily from the fact the columns of A are linearly independent. ∎

6 Ordinary Least Squares and its Application in Statistics

6.1 The Orthogonal Decomposition Theorem and Least-Squares Solution

Theorem 6.1. The Orthogonal Decomposition Theorem: Let W be a subspace of $\mathbb{R}^n$. Then, each $\mathbf{y}$ in $\mathbb{R}^n$ can be written **uniquely** in the form

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{z} \tag{6.1}$$

where $\hat{\mathbf{y}}$ is in W and $\mathbf{z}$ is in $W^{\perp}$. In fact, if $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p\}$ is any *orthogonal basis* of W, then

$$\hat{\mathbf{y}} = \frac{\mathbf{y}^\top \mathbf{u}_1}{\mathbf{u}_1^\top \mathbf{u}_1} + \frac{\mathbf{y}^\top \mathbf{u}_2}{\mathbf{u}_2^\top \mathbf{u}_2} + \cdots + \frac{\mathbf{y}^\top \mathbf{u}_p}{\mathbf{u}_p^\top \mathbf{u}_p} \tag{6.2}$$

and $\mathbf{z} = \mathbf{y} - \hat{\mathbf{y}}$.

Definition 6.1. If X is $n \times p$ and $\boldsymbol{\beta}$ is in $\mathbb{R}^p$, a **least-squares solution** of of $X\boldsymbol{\beta} = \mathbf{y}$ is an $\hat{\boldsymbol{\beta}}$ in $\mathbb{R}^p$ such that

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p : \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\| \leq \|\mathbf{y} - X\boldsymbol{\beta}\| \tag{6.3}$$

We cannot ensure that the linear system $X\boldsymbol{\beta} = \mathbf{y}$ is always consistent. That is, $\mathbf{y}$ may not be in $\mathrm{Col}(X)$. But we can find a $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that equation (6.3) holds. Let $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$, by **Orthogonal Decomposition Theorem** $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to $\mathrm{Col}(X)$, this is,

$$\forall i \in \{1, 2, \cdots, p\} : \mathbf{x}_i^\top (\mathbf{y} - \hat{\mathbf{y}}) = 0$$

and thus,

$$X^\top (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0} \Longrightarrow X^\top (\mathbf{y} - X\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

We can find $\hat{\boldsymbol{\beta}}$ by solving the following linear system, which is called **normal equation** and must be consistent

$$X^\top \mathbf{y} = X^\top X \hat{\boldsymbol{\beta}} \tag{6.4}$$

Furthermore, if $(X^\top X)^{-1}$ exists,

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y} \tag{6.5}$$

The same result can be derived from example 13.1, using vector calculus. The prediction vector $\hat{\mathbf{y}}$ (the projection of $\mathbf{y}$ onto $\mathrm{Col}(X)$) can thus be expressed as

$$\hat{\mathbf{y}} = X(X^\top X)^{-1} X^\top \mathbf{y} = X\hat{\boldsymbol{\beta}} \tag{6.6}$$

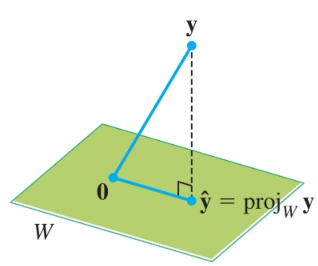


Figure 2: $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto W, where W is the column space of $\mathbf{X}$.

Remark 6.1. In linear model, we are interested in the difference of the response vector $\mathbf{y}$ and its projection onto the column space of design matrix X. The **Sum of Squares due error** is a measurement for that purpose, which is defined as:

$$\mathrm{SSE}(\mathbf{y}) = \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2 \tag{6.7}$$

6.2 Orthogonal Projection Matrix

By equation (6.6), we can see that the effect of $X(X^\top X)^{-1}X^\top$ is to project $\mathbf{y}$ onto $\mathrm{Col}(X)$, which is why it is called **(Orthogonal) Projection Matrix**. The projection matrix is also called **Hat Matrix** in statistics. The hat matrix differs from equation (5.4), which projects a vector onto a vector, **while the hat matrix projects a vector onto the column space of X**.

Remark 6.2. The hat matrix is typically denoted by H, it has the following properties:

1. H is symmetric and thus a square matrix.
2. $H^2 = H$.
3. If $\mathbf{x} \in \mathrm{Col}(X)$, $H\mathbf{x} = \mathbf{x}$.

Definition 6.2. Idempotent Matrix: A square matrix A is said to be idempotent if and only if $A^2 = A$.

Definition 6.3. Orthogonal Projection Matrix: A matrix P is an orthogonal projection matrix if P is **idempotent and symmetric**.

Remark 6.3. For any vector $\mathbf{y}$, P projects $\mathbf{y}$ onto a subspace W, resulting in $\hat{\mathbf{y}} = P\mathbf{y}$. If we project $\hat{\mathbf{y}}$ onto W again, the equation

$$\hat{\mathbf{y}} = P\hat{\mathbf{y}} = PP\mathbf{y} = P^2\mathbf{y} \tag{6.8}$$

illustrated why P is needed to be **idempotent**. Conversely, suppose we want to project a vector $\mathbf{y}$ onto a subspace W spanned by $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$, we can find an orthonormal base by Gram-Schmidt Process, say $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$. By theorem 5.4,

$$\hat{\mathbf{y}} = UU^\top\mathbf{y}$$

If we let $P = UU^\top$, P is clearly **symmetric**. Note also that **P projects a vector onto the subspace spanned by the columns (or rows, since P is symmetric) of P**.

We have already known that the hat matrix H is the orthogonal projection matrix onto the column space of X, and the residual vector

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - H\mathbf{y}) = (I - H)\mathbf{y}$$

is orthogonal to $\mathrm{Col}(X)$. It is intuitive to say that $I - H$ is the orthogonal projection matrix onto $\mathrm{Col}(X)^\perp$ or $\mathrm{Nul}(X^\top)$.

Theorem 6.2. If P is an orthogonal projection matrix, then $I - P$ is an orthogonal projection matrix onto $\mathrm{Col}(P)^\perp$ (or $\mathrm{Row}(P)^\perp$, since P is symmetric.

Theorem 6.3. The eigenvalues of an orthogonal projection matrix P are either 1's or 0's.

Proof. Since $\forall \mathbf{x} \in \mathrm{Col}(P)$: $H\mathbf{x} = \mathbf{x}$, $\mathrm{Col}(P)$ is an eigenspace of P corresponding to the eigenvalue 1. And $\forall \mathbf{v} \in \mathrm{Col}(P)^\perp : P\mathbf{v} = 0 \cdot \mathbf{v}$ says that $\mathrm{Col}(P)^\perp$ is another eigenspace of P corresponding to eigenvalue 0. P is a $n \times n$ symmetric matrix, and $\dim\Big(\mathrm{Col}(P)\Big) + \dim\Big(\mathrm{Col}(P)^\perp\Big) = n$, so by theorem 9.11 H can only have eigenvalues of 0 or 1. □

Theorem 6.4. An orthogonal projection matrix is semi-positive definite.

Proof. By theorem 6.3 and theorem 10.2. □

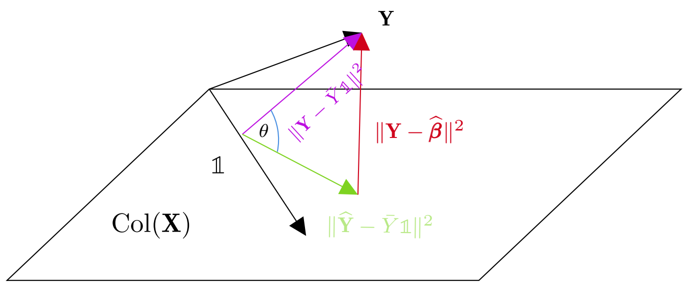


Figure 3: SST, SSE and SSR form a right triangle.

Example 6.1. Let a quadratic form $Q(\mathbf{x}) = \mathbf{x}^\top (I - H_0)\mathbf{x}$, where H_0 is the projection matrix onto $\mathbb{1}$ as discussed in equation (5.5). Given a vector, what does $Q(\mathbf{x})$ stand for?

$$\begin{aligned}
Q(\mathbf{x}) &= \mathbf{x}^\top (I - H_0)\mathbf{x} \\
&= \|\mathbf{x}\|^2 - \mathbf{x}^\top H_0 \mathbf{x} \\
&= \|\mathbf{x}\|^2 - \mathbf{x}^\top \bar{x}\mathbb{1} \\
&= \|\mathbf{x}\|^2 - \bar{x}\sum_{i=1}^{n} x_i \\
&= \|\mathbf{x}\|^2 - n\bar{x}^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1)S^2
\end{aligned}$$

where S^2 is the sample variance.

6.3 Application in Linear Model

We have calculated SST by equation (5.6), and we want to calculate the projection vector of $\mathbf{y} - H_0\mathbf{y}$ onto Col(X)

$$H(\mathbf{y} - H_0\mathbf{y}) = H\mathbf{y} - HH_0\mathbf{y} = X\hat{\boldsymbol{\beta}} - \bar{\mathbf{y}}\ \mathbb{1}$$

where $H\mathbf{y}$ is the prediction vector by equation (6.6) and $HH_0\mathbf{y} = H_0\mathbf{y}$ since $H_0\mathbf{y}$ is in Col(X). That is so-called **Sum of Squares due to Regression**, which is defined as:

$$\mathrm{SSR}(\mathbf{y}) = \|X\hat{\boldsymbol{\beta}} - \bar{y}\ \mathbb{1}\|^2 = \sum_{i=1}^{n}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - \bar{y})^2 \tag{6.9}$$

Theorem 6.5. We have calculated SST, SSR and SSE, there is a relationship between them:

$$\mathrm{SST}(\mathbf{y}) = \mathrm{SSR}(\mathbf{y}) + \mathrm{SSE}(\mathbf{y}) \tag{6.10}$$

Or equivalently,

$$\|\mathbf{y} - \bar{y}\ \mathbb{1}\|^2 = \|X\hat{\boldsymbol{\beta}} - \bar{y}\ \mathbb{1}\|^2 + \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2 \tag{6.11}$$

Proof. It can be proved by Pythagorean Theorem as shown in figure 3. □

Theorem 6.6. Suppose V is subspace of $\mathbb{R}^p$, and W is a subspace of V, that is, $W \subseteq V$ and $\dim(W) \leq \dim(V)$. Then

$$\forall \mathbf{y} \in \mathbb{R}^p : \|\text{proj}_V(\mathbf{y})\| \geq \|\text{proj}_W(\mathbf{y})\| \tag{6.12}$$

Proof. By theorem 6.1, $\mathbf{y} = \text{proj}_W(\mathbf{y}) + \mathbf{r}_W = \text{proj}_V(\mathbf{y}) + \mathbf{r}_V$. Since $W \subseteq V$, $\text{proj}_V(\mathbf{y}) - \text{proj}_W(\mathbf{y}) \in V$. We can draw a triangle with $\mathbf{r}_W$ as the hypotenuse, $\mathbf{r}_V$ and $\text{proj}_V(\mathbf{y}) - \text{proj}_W(\mathbf{y})$ as the legs. we have

$$\|\mathbf{r}_W\| > \|\mathbf{r}_V\| + \|\text{proj}_V(\mathbf{y}) - \text{proj}_W(\mathbf{y})\| \tag{6.13}$$

It indicates that $\|\mathbf{r}_W\| > \|\mathbf{r}_V\|$. By Pythagorean Theorem

$$\|\mathbf{y}\|^2 = \|\text{proj}_W(\mathbf{y})\|^2 + \|\mathbf{r}_W\|^2 = \|\text{proj}_V(\mathbf{y})\|^2 + \|\mathbf{r}_V\|^2$$

Thus, $\|\text{proj}_V(\mathbf{y})\| > \|\text{proj}_W(\mathbf{y})\|$. Note that $\|\text{proj}_V(\mathbf{y})\| = \|\text{proj}_W(\mathbf{y})\|$ if and only if $V = W$. ☐

The theorem 6.6 provides **an interesting insight for the design matrix X. If we add a new column (or a new feature) into X, resulting in a new matrix $\tilde{X}$, then SSE($\mathbf{y}$) would not increase.** Looking at equation (6.11), $\|\mathbf{y} - \overline{y}\ \mathbb{1}\|^2$ is a constant and $\|\mathbf{r}_V\| = \|\mathbf{y} - \tilde{X}\hat{\boldsymbol{\beta}}\| \leq \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\| = \|\mathbf{r}_W\|$ as in equation (6.13), since $\dim(X) \leq \dim(\tilde{X})$. Therefore, **in no case will the *SSE* increase, because the model now has more capacity to minimize the residuals (or in other words, it has more freedom to find a better fit).**

The three vectors, $\mathbf{y} - \bar{y} - \mathbb{1}$, $X\hat{\boldsymbol{\beta}} - \bar{y}\ \mathbb{1}$ and $\mathbf{y} - X\hat{\boldsymbol{\beta}}$, forms a right triangle. We can use the cosine value, as shown in figure 3, to reflect the length of $\mathbf{y} - X\hat{\boldsymbol{\beta}}$:

$$\cos^2\theta = \frac{\text{SSR}(\mathbf{y})}{\text{SST}(\mathbf{y})}$$

We can see that the range of $\cos^2\theta$ is $[0, 1]$, and its value is proportional to SSR($\mathbf{y}$).

Definition 6.4. The coefficient of determination:

$$R^2 = 1 - \frac{\text{SSE}(\mathbf{y})}{\text{SST}(\mathbf{y})} = \frac{\text{SSR}(\mathbf{y})}{\text{SST}(\mathbf{y})} \tag{6.14}$$

The higher the R^2 is, the more accurate the predictions of our model are. R^2 **non-decreases (by theorem 6.6)** as we add new features (or columns) into the design matrix X.

7 Data Projection

Consider the following matrix multiplication

$$Z = XV$$

where $X = (\mathbf{x}_1 \ \cdots \ \mathbf{x}_n)^\top$ is $n \times p$ and V is $p \times p$.

$$Z = XV = \begin{bmatrix} \mathbf{x}_1^\top V \\ \vdots \\ \mathbf{x}_n^\top V \end{bmatrix} = \begin{bmatrix} \mathbf{z}^{(1)} \\ \vdots \\ \mathbf{z}^{(n)} \end{bmatrix} \tag{7.1}$$

$\mathbf{z}^{(i)} = \mathbf{x}_i^\top V$ can be considered as a linear combination of rows of V using the entries in $\mathbf{x}_i^\top$ as weights. This implies

$$\mathbf{x}_i = V^\top (\mathbf{z}^{(i)})^\top$$

$(\mathbf{z}^{(i)})^\top$ is the coordinate of $\mathbf{x}_i$ relative to the rows of V. Furthermore, the j^{th} entry, $z_{ij} = \mathbf{x}_i^\top \mathbf{v}_j$, in $(\mathbf{z}^{(i)})^\top$ is the scalar projection of $\mathbf{x}_i^\top$ on $\mathbf{v}_j$ or on span($\mathbf{v}_j$). Looking at (11), let $X_j = X\mathbf{v}_j \otimes \mathbf{v}_j$ and $\mathbf{x}_j^{(i)}$ be the i^{th} row of X_j, then

$$\mathbf{x}_j^{(i)} = \mathbf{x}_i^\top \mathbf{v}_j \mathbf{v}_j^\top = z_{ij} \mathbf{v}_j^\top$$

which is the projection vector of $\mathbf{x}_i^\top$ on $\mathbf{v}_j$. That is, **the rows of X_j are the vector projections of rows of X on $\mathbf{v}_j^\top$.**

Since all the rows of X_j are the projections on $\mathbf{v}_j^\top$, we have

$$\text{rank}(\mathbf{v}_j \otimes \mathbf{v}_j) = 1 \Longrightarrow \text{rank}(X_j) = 1$$

All data points (or rows) of X_j are on the line that goes through the origin and vector $\mathbf{v}_j^\top$. It says that we can restore XV to X by right-multiplying it by $V^\top$

$$\begin{aligned} XVV^\top &= X\mathbf{v}_1 \otimes \mathbf{v}_1 + \cdots + X\mathbf{v}_n \otimes \mathbf{v}_n \\ &= X_1 + \cdots + X_n \\ &= X \end{aligned}$$

Again, each row of XV represents the coordinate of $(\mathbf{v}_1 \ \cdots \ \mathbf{v}_p)^\top$. By right-multiplying it by its inverse $V^\top$, we can restore the coordinates to those of **standard orthonormal basis**. Another way to view $XVV^\top$ is as the sum of the projections of all data points onto the orthonormal basis.

8 Rank and Trace

8.1 Rank

Definition 8.1. The **rank** of a matrix $A \in \mathbb{R}^{n\times p}$ is the number of its linearly independent columns (or rows), which is expressed as rank(A).

Given a matrix $A \in \mathbb{R}^{n\times p}$, it has the following properties:

1. $\text{rank}(A) = \min\{n, p\}$
2. $\text{rank}(AB) = \min\{\text{rank}(A), \text{rank}(B)\}$
3. Given two non-singular matrices $B \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times p}$:

$$\text{rank}(BA) = \text{rank}(AC) = \text{rank}(A) \tag{8.1}$$

4. $\text{rank}(A^\top A) = \text{rank}(AA^\top) = \text{rank}(A) = \text{rank}(A^\top)$

Note that: property 4 illustrates that **multiplying A by a non-singular matrix does not change the rank of** A.

Example 8.1. Show that if a matrix $A \in \mathbb{R}^{n\times p}$ with $n \geq p$ is of full column rank, then $A^\top A$ is non-singular.

Proof. Since A is of full column rank and $n \geq p$, we have

$$\text{rank}(A) = p = \text{rank}(A^\top A)$$

Since $A^\top A$ is a $p \times p$ matrix and has full column rank, it is non-singular. □

Example 8.2. Show that if a matrix $A \in \mathbb{R}^{n\times p}$ with $n \geq p$ is not of full column rank, then $A^\top A$ is singular.

Proof. Since A is not of full column rank,

$$\text{rank}(A) = \text{rank}(A^\top A) < p$$

It implies that A is singular. □

Example 8.3. Show that given a matrix $A \in \mathbb{R}^{n\times p}$ with $n < p$, $A^\top A$ is singular.

Proof. Since $\text{rank}(A) \leq \min\{n, p\}$,

$$\text{rank}(A) = \text{rank}(A^\top A) \leq n < p$$

Since $A^\top A$ is not of full column rank, it is singular. □

8.2 Trace

Definition 8.2. The trace of a square matrix $A \in \mathbb{R}^{n\times n}$ is the sum of diagonal elements of A. It is denoted $\text{tr}(A) = \sum_{i=1}^{n} = a_{ii}$.

Theorem 8.1. The trace function $\text{tr}(\cdot)$ has the following properties:

1. $\text{tr}(cA \pm dB) = c\text{tr}(A) \pm d\text{tr}(B)$, where $c, d \in \mathbb{R}$.
2. Given two matrices $A \in \mathbb{R}^{n\times p}, B \in \mathbb{R}^{p\times n}$, then $\text{tr}(AB) = \text{tr}(BA)$

   Proof. Let t_i be the i^{th} elements on the diagonal of AB. Then

$$\text{tr}(AB) = \sum_{i=1}^{n} t_i = \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}b_{ji} = \sum_{j=1}^{p}\sum_{i=1}^{n} b_{ji}a_{ij} = \text{tr}(BA)$$

   Note that n is not required to be greater or equal to p. □

3. Given an $n \times p$ matrix, $A = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_p\end{bmatrix}$, $\text{tr}(A^\top A) = \sum_{i=1}^{p} \mathbf{a}_i^\top \mathbf{a}_i$
4. Given an $n \times p$ matrix, $\text{tr}(AA^\top) = \sum_{i=1}^{n} \mathbf{a}^{(i)}\mathbf{a}_i$, where $\mathbf{a}^{(1)}$ is the row vector of A.
5. By property 3 and 4, $\text{tr}(A^\top A) = \text{tr}(AA^\top) = \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}^2$
6. $\text{tr}(\mathbb{E}(\mathbf{X})) = \mathbb{E}(\text{tr}(\mathbf{X}))$, where $\mathbb{E}$ represents the expectation of a random matrix.

9 Eigenvalues and Diagonalization

9.1 Eigenvectors

Definition 9.1. Given a square matrix $A \in \mathbb{R}^{n\times n}$, there exists a vector $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that

$$\exists \lambda \in \mathbb{R} : \ A\mathbf{x} = \lambda \mathbf{x} \tag{9.1}$$

where λ is called an **eigenvalue** of A; $\mathbf{x}$ is called an **eigenvector corresponding to** λ.

The equation (9.1) can be rewritten as

$$(A - \lambda I)\mathbf{x} = \mathbf{0} \tag{9.2}$$

This implies that **the set of all solutions of equation (9.2)** is just the null space $\text{Nul}(A - \lambda I)$. So this set is a *subspace* of $\mathbb{R}^n$ and its called the **eigenspace** of A corresponding to λ.

Definition 9.2. A scalar λ is an eigenvalue of a matrix $A \in \mathbb{R}^n$ if and only if λ satisfies the **characteristic equation**:

$$\det(A - \lambda I) = 0 \tag{9.3}$$

Remark 9.1. $A\mathbf{x} = 0\mathbf{x}$ holds if and only if A is singular. That is, 0 **is an eigenvalue of** A **in and only if** A **is singular**.

Theorem 9.1. The eigenvalues of a **triangular matrix** are the entries on its main diagonal.

Theorem 9.2. If $\mathbf{v}_1, \cdots, \mathbf{v}_r$ are eigenvalues that correspond to distinct eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_r$, then the set $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ is **linearly independent**.

9.2 Similarity

Definition 9.3. If A and B are $n \times n$ matrices, then A is similar to B if there is a non-singular matrix P such that

$$P^{-1}AP = B$$

Theorem 9.3. Given two matrices $A, B \in \mathbb{R}^{n\times n}$, if A and B are similar, then they have the same characteristic polynomial and hence the same eigenvalues(with the same multiplicities).

Remark 9.2. However, A and B having the exactly same eigenvalues does not imply that A and B are similar.

9.3 Diagonalization

In many cases, the eigenvalue-eigenvector information contained within a matrix A can be displayed in a useful factorization for the form $A = PDP^{-1}$ where D is a diagonal matrix.

Theorem 9.4. The Diagonalization Theorem: Given a matrix $A \in \mathbb{R}^{n\times n}$, A is diagonalizable if and only if A has n linearly independent eigenvectors.

Remark 9.3. In fact $A = PDP^{-1}$, if and only if the columns of P are n **linearly independent** eigenvectors of A. In this case, the diagonal entries of D are eigenvalues of A that correspond, resprectively, to the eigenvectors in P.

In other words, A is diagonalizable if and only if there are enough eigenvectors to form a basis of $\mathbb{R}^n$, which is called an **eigenvector basis** of $\mathbb{R}^n$.

Theorem 9.5. An $n \times n$ mateix with n **distinct eigenvalues** is diagonalizable.

Theorem 9.6. Let A be an $n \times n$ matrix whose distinct eigenvalues are $\lambda_1, \cdots, \lambda_p$. Let $\dim(\mathcal{E}(\lambda_k))$ denote the dimension of eigenspace for λ_k. The matrix A is diagonalizable if and only if

$$\sum_{i=1}^{p} \dim(\mathcal{E}(\lambda_i)) = n$$

Theorem 9.7. If A with p distinct eigenvalues is diagonalizable and $\mathcal{B}_k$ is a basis for the eigenspace corresponding to λ_k, then the total collection of vectors in the sets $\mathcal{B}_1, \cdots, \mathcal{B}_p$ forms an eigenvector basis in $\mathbb{R}^n$.

9.4 Eigenvectors and Linear Transformation

We have already understood the simple linear transformation $A\mathbf{x}$. The goal of this section is to understand the nested transformation of $A = PDP^{-1}$.

Definition 9.4. Standard Matrix: Any Linear transformation $T : \mathbb{R}^p \mapsto \mathbb{R}^n$ can be implemented via left-multiplication by a matrix A, called the **standard matrix** of T.

Let V be a p-dimensional vector space, let W be an n-dimensional vector space, and let T be any linear transformation from V to W. To associate a matrix with T, choose ordered bases $\mathcal{B}$ and $\mathcal{C}$ for V and W, respectively.

$\forall \mathbf{x} \in V$, the coordinate vector $[\mathbf{x}]_\mathcal{B}$ is in $\mathbb{R}^p$, and the coordinate vector of its image, $[T(\mathbf{x})]_\mathcal{C}$ is in $\mathbb{R}^n$. if $\mathbf{x} = r_1\mathbf{b}_1 + r_1\mathbf{b}_2 + \cdots + r_1\mathbf{b}_p$, then

$$[\mathbf{x}]_\mathcal{B} = \begin{bmatrix} r_1 \\ \vdots \\ r_p \end{bmatrix}$$

and

$$T(\mathbf{x}) = T(r_1\mathbf{b}_1 + r_1\mathbf{b}_2 + \cdots + r_1\mathbf{b}_p) = r_1T(\mathbf{b}_1) + r_2T(\mathbf{b}_2) + \cdots + r_pT(\mathbf{b}_p) \tag{9.4}$$

Since the coordinate mapping from W to $\mathbb{R}^n$ is linear, equation (9.4) leads to

$$[T(\mathbf{x})]_\mathcal{C} = r_1[T(\mathbf{b}_1)]_\mathcal{C} + r_2[T(\mathbf{b}_2)]_\mathcal{C} + \cdots + r_p[T(\mathbf{b}_p)]_\mathcal{C} \tag{9.5}$$

Since $\mathcal{C}$-coordinate vecotrs are in $\mathbb{R}^n$, the vector equation (9.5) can be written as a matrix equation, namely,

$$[T(\mathbf{x})]_\mathcal{C} = M[\mathbf{x}]_\mathcal{B} \tag{9.6}$$

where

$$M = \begin{bmatrix} T(\mathbf{b}_1)]_\mathcal{C} & T(\mathbf{b}_2)]_\mathcal{C} & \cdots & T(\mathbf{b}_p)]_\mathcal{C} \end{bmatrix}$$

The matrix M is a matrix representation of T, called the **matrix for T relative to the bases $\mathcal{B}$ and $\mathcal{C}$**. In the common case where W is the same as V and the basis $\mathcal{C}$ is the same as $\mathcal{B}$, the matrix M in equation (9.6) is called the **matrix for T relative to $\mathcal{B}$**, or simply the **$\mathcal{B}$-matrix for T**, and is denoted by $[T]_\mathcal{B}$. The $\mathcal{B}$-matrix for $T : V \to V$ satisfies:

$$[T(\mathbf{x})]_\mathcal{B} = [T]_\mathcal{B}[\mathbf{x}]_\mathcal{B}$$

Theorem 9.8. Diagonal Matrix Representation: Suppose $A = PDP^{-1}$, where D is a diagonal $n \times n$ matrix. If $\mathcal{B}$ is the basis for $\mathbb{R}^n$ formed from the columns of P, then D is the $\mathcal{B}$-matrix for the transformation.

Proof. Let $\mathcal{B} = \{\mathbf{b}_1, \cdots, \mathbf{b}_n\}$ and $P = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{bmatrix}$. In this case, P is the change-of-coordinates matrix $P_\mathcal{B}$ discussed in definition 4.2, where

$$P[\mathbf{x}]_\mathcal{B} = \mathbf{x} \quad \text{and} \quad [\mathbf{x}]_\mathcal{B} = P^{-1}\mathbf{x}$$

If $T(\mathbf{x}) = A\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$, then

$$\begin{aligned} [T]_\mathcal{B} &= \begin{bmatrix} [T(\mathbf{b}_1)]_\mathcal{B} & \cdots & [T(\mathbf{b}_n)]_\mathcal{B} \end{bmatrix} \\ &= \begin{bmatrix} [A\mathbf{b}_1]_\mathcal{B} & \cdots & [A\mathbf{b}_n]_\mathcal{B} \end{bmatrix} \\ &= \begin{bmatrix} P^{-1}A\mathbf{b}_1 & \cdots & P^{-1}A\mathbf{b}_n \end{bmatrix} \\ &= P^{-1}A \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{bmatrix} \\ &= P^{-1}AP = D \end{aligned}$$

□

Remark 9.4. The proof of theorem 9.8 didn't use the information that D was diagonal. Hence, if A is similar to a matrix C, with $A = PCP^{-1}$, then C is the $\mathcal{B}$-matrix for the transformation $\mathbf{x} \mapsto A\mathbf{x}$ when the basis $\mathcal{B}$ is formed from the columns of P. Multiplying by such a matrix A has the following interpretation: given a vector $\mathbf{x} \in V$

1. $P^{-1}\mathbf{x} \mapsto [\mathbf{x}]_\mathcal{B}$
2. $C[\mathbf{x}]_\mathcal{B} \mapsto [A\mathbf{x}]_\mathcal{B}$
3. $P[A\mathbf{x}]_\mathcal{B} \mapsto A\mathbf{x}$

$$\begin{tikzcd} \mathbf{x} \arrow[r, "A"'] \arrow[d, "P^{-1}"'] & A\mathbf{x} \\ {[\mathbf{x}]_\mathcal{B}} \arrow[r, "C"'] & {[A\mathbf{x}]_\mathcal{B}} \arrow[u, "P"'] \end{tikzcd}$$

Conversely, if $T : \mathbb{R}^n \to \mathbb{R}^n$ is defined by $T(\mathbf{x}) = A\mathbf{x}$, and if $\mathcal{B}$ is any basis for $\mathbb{R}^n$, then the $\mathcal{B}$-matrix for T is similar to A. The theorem 9.8 show that if P is the matrix whose columns come from the vectors in $\mathcal{B}$, then

$$[T]_{\mathcal{B}} = P^{-1}AP$$

Thus, the set of all matrices similar to a matrix A **coincides with the set of all matrix representations of the transformation $\mathbf{x} \mapsto A\mathbf{x}$.**

9.5 Symmetric Matrices

Definition 9.5. A **symmetric** matrix is a matrix A such that $A^\top = A$. Note that such a matrix is necessarily square.

Theorem 9.9. If A is symmetric, then any two eigenvectors from different eigenspaces are orthogonal.

Proof. Suppose there are two eigenvectors $\mathbf{v}_1, \mathbf{v}_2$, respectively, corresponding to distinct eigenvalues λ_1 and λ_2. Consider the following equation:

$$\begin{aligned}
\lambda_1 \mathbf{v}_1^\top \mathbf{v}_2 &= (A\mathbf{v}_1)^\top \mathbf{v}_2 \\
&= \mathbf{v}_1^\top A^\top \mathbf{v}_2 \\
&= \mathbf{v}_1^\top A \mathbf{v}_2 \quad \text{since } A \text{ is a symmetric matrix} \\
&= \lambda_2 \mathbf{v}_1^\top \mathbf{v}_2
\end{aligned}$$

We can get $(\lambda_1 - \lambda_2)\mathbf{v}_1^\top \mathbf{v}_2 = 0$. Since $\lambda_1 \neq \lambda_2$, $\mathbf{v}_1^\top \mathbf{v}_2$ must be 0. □

Definition 9.6. Orthogonally dianonalizable: For an $n \times n$ matrix A, if there are an **orthogonal matrix** P with $(P^{-1} = P^\top)$ and a diagonal matrix D such that

$$A = PDP^\top = PDP^{-1} \tag{9.7}$$

then A is said to be **Orthogonally dianonalizable**.

Remark 9.5. Such a diagonalization requires n **linearly independent** and **orthonormal eigenvectors**. If A is orthogonally diagonalizable as in equation (9.7), then

$$A^\top = (PDP^\top)^\top = PDP^\top = A$$

Thus, A is symmetric.

Theorem 9.10. An $n \times n$ matrix A is orthogonally diagonalizable if and only if A is a symmetric matrix.

Theorem 9.11. The Spectral Theorem for Symmetric Matrices: An $n \times n$ matrix A has the following properties:

1. A has n real eigenvalues, counting multiplicities.
2. The dimension of the eigenspace for each eigenvalue λ equals the multiplicity of λ as a root of the characteristic equation.
3. The eigenspaces are mutually orthogonal, in the sense that eigenvectors corresponding to different eigenvalues are orthogonal.
4. A is orthogonally diagonalizable.

Theorem 9.12. Spectral Decomposition: Suppose A is orthogonally diagonalizable,

$$
\begin{aligned}
A =& PDP^\top = \begin{bmatrix}[\mathbf{u}_1 & \cdots & \mathbf{u}_n]\end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^\top \\ \vdots \\ \mathbf{u}_n^\top \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 \mathbf{u}_1 & \cdots & \lambda_n \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^\top \\ \vdots \\ \mathbf{u}_n^\top \end{bmatrix}
\end{aligned}
$$

Using the equation (2.2), the sum of outer product representation:

$$
A = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^\top + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^\top + \cdots + \lambda_n \mathbf{u}_n \mathbf{u}_n^\top \tag{9.8}
$$

This representation of A is called a **spectral decomposition** of A. Note each $\mathbf{u}_i \mathbf{u}_1 i\top$ is a projection matrix with rank 1.

9.6 Intuition of Unit Eigenvectors

Suppose that a symmetric matrix $A \in \mathbb{R}^2$ with two *unit* eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$, which are orthogonal to each other.

1. We can find a unit circle that goes through the four points: $\mathbf{v}_1, \mathbf{v}_2, -\mathbf{v}_1, -\mathbf{v}_2$. After multiplying the four vectors by A, we can find an ellipse that goes through these fore vectors.
2. Suppose $A \in \mathbb{R}^{n\times n}$ can be diagonalized into

   $$A = PDP^{-1}$$

   Right-multiplying A by P:

   $$AP = \begin{bmatrix} A\mathbf{v}_1 & A\mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{v}_1 & \lambda_2 \mathbf{v}_2 \end{bmatrix}$$

   We can find an ellipse that goes through the columns of AP. Actually P is an orthogonal matrix, its effect is to perform **a rotational transformation**, mapping a coordinate vector relative to $\{\mathbf{v}_1, \mathbf{v}_2\}$ to a vector in the standard basis. While the effect of the diagonal matrix D is to perform **a scaling transformation**.
3. Consider the linear transformation $T(\mathbf{x}) = A\mathbf{x}$,

   $$A\mathbf{x} = PDP^{-1}\mathbf{x} = \mathbf{y}$$

   $P^{-1}\mathbf{x} = [\mathbf{x}]_\mathcal{B}$ maps $\mathbf{x}$ into a new coordinate system with a set of orthonormal basis as its coordinate vectors, which corresponds to a **rotational action**. $D[\mathbf{x}]_\mathcal{B} = [\mathbf{y}]_\mathcal{B}$ scales the vector. $P[\mathbf{y}]_\mathcal{B}$ transforms $[\mathbf{y}]_\mathcal{B}$ back to standard basis.
4. Multiplying A by a vector or a matrix (a set of column vectors) corresponds to a sequence of operations: a rotation, followed by a scaling, and then a rotation back.

9.7 Important Properties of Eigenvalues

If $\mathbf{v}$ is an eigenvector of A corresponding to eigenvalue λ, then

$$A^2\mathbf{v} = AA\mathbf{v} = A\lambda\mathbf{v} = \lambda^2\mathbf{v}$$

We can generalize the equation above to

$$A^k\mathbf{v} = \lambda^k\mathbf{v} \tag{9.9}$$

Suppose $A \in \mathbb{R}^{n\times n}$ has n eigenvalues, then

$$\det(A) = \prod_{i=1}^{n} \lambda_i \tag{9.10}$$

Proof. Let the characteristic equation of A be

$$p(\lambda) = \det(A - \lambda I) = (\lambda_1 - \lambda)\cdots(\lambda_n - \lambda)$$

We can simply get the result by letting λ be zero. □

9.8 Spectral Decomposition on Gram Matrix

Given a data matrix $X \in \mathbb{R}^{n\times p}$, its Gram matrix $X^\top X$ is symmetric and, therefore, orthogonally diagonalizable by theorem 9.10:

$$G = X^\top X = PDP^\top$$

We can get the following equation:

$$P^\top GP = \begin{bmatrix} \mathbf{u}_1^\top X^\top X\mathbf{u}_1 & \mathbf{u}_1^\top X^\top X\mathbf{u}_2 & \cdots & \mathbf{u}_1^\top X^\top X\mathbf{u}_p \\ \dots & \dots & \ddots & \dots \\ \mathbf{u}_p^\top X^\top X\mathbf{u}_1 & \mathbf{u}_p^\top X^\top X\mathbf{u}_2 & \cdots & \mathbf{u}_p^\top X^\top X\mathbf{u}_p \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Let $X\mathbf{u}_i = \mathbf{y}_i$, the k^{th} entry of $\mathbf{y}_i$ is the projection of the k^{th} data point (the k^{th} row of X) onto the eigenvector $\mathbf{u}_i$.

$$P^\top GP = \begin{bmatrix} \mathbf{y}_1^\top \mathbf{y}_1 & \mathbf{y}_1^\top \mathbf{y}_2 & \cdots & \mathbf{y}_1^\top \mathbf{y}_p \\ \dots & \dots & \ddots & \dots \\ \mathbf{y}_p^\top \mathbf{y}_1 & \mathbf{y}_p^\top \mathbf{y}_2 & \cdots & \mathbf{y}_p^\top \mathbf{y}_p \end{bmatrix} = D$$

For any $i \neq j$, we can see that $\mathbf{y}_i$ and $\mathbf{y}_j$ are orthogonal to each other. Meanwhile, $\|\mathbf{y_i}\|^2 = \lambda_i$, which means

$$\sum_{j=1}^{p} y_{ij}^2 = \lambda_i$$

That is, the sum of squares of the coordinates of each data point relative to $\mathbf{y}_i$ equals to λ_i. This means that the projections of data points (rows) of X onto different eigenvectors of G have different sums of squares. We can express G in its spectral decomposition form:

$$G = \lambda_1\mathbf{u}_1\mathbf{u}_1^\top + \lambda_2\mathbf{u}_2\mathbf{u}_2^\top + \cdots + \lambda_n\mathbf{u}_n\mathbf{u}_n^\top \tag{9.11}$$

The equation above indicates that **the larger the eigenvalue, the more important the eigenvector**, as the projections of data points onto it are larger.

9.9 Change of Variable

Suppose $A \in \mathbb{R}^{n\times n}$ has n eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$, which can form a basis $\mathcal{B}$ for $\mathbb{R}^n$. Let $\begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_n\end{bmatrix}$. Given a sequence $\{\mathbf{x}_k\}$ satisfying

$$\mathbf{x}_{k+1} = A\mathbf{x}_k$$

which is called a difference equation. Define a new sequence $\{\mathbf{y}_k\}$ by

$$\mathbf{y}_k = P^{-1}\mathbf{x}_k, \quad \text{or equivalently,} \quad \mathbf{x}_k = P\mathbf{y}_k$$

$\mathbf{y}_k$ is clearly the coordinate of $\mathbf{x}_k$ relative to $\mathcal{B}$ by definition 4.2. Substituting these relations into the equation $\mathbf{x}_{k+1} = A\mathbf{x}_k$ and using the fact that $A = PDP^{-1}$:

$$\mathbf{x}_{k+1} = AP\mathbf{y}_k = PDP^{-1}P\mathbf{y}_k = PD\mathbf{y}_k$$

Left-multiplying the above equation by P^{-1}:

$$P^{-1}\mathbf{x}_{k+1} = \mathbf{y}_{k+1} = D\mathbf{y}_k$$

The change of variable from $\mathbf{x}_k$ to $\mathbf{y}_k$ has **decoupled** the system of difference equations. Geometrically, the only effect on $\mathbf{y}_k$ is scaling the vector, and each entry y_i of $\mathbf{y}_k$ is unaffected by the other entries. **Decoupling the system allows for the calculation in a new coordinate system, which demonstrates the power of linear algebra.**

10 TODO: Quadratic Form

Definition 10.1. A **quadratic form on** $\mathbb{R}^n$ is a function $Q : \mathbb{R}^n \to \mathbb{R}$ whose input vector $\mathbf{x}$ can be computed by an expression of the form:.

$$Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$$

where A is a $n \times n$ *symmetric matrix* and called **the matrix of the quadratic form**. Since A is symmetric, $Q(\mathbf{x})$ can also be expressed as:

$$Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij}x_ix_j$$

10.1 Change of Variable in a Quadratic Form

Let $\mathbf{x} \in \mathbb{R}^n$, then a *change of variable* is an equation of the form

$$\mathbf{x} = P\mathbf{y}, \quad \text{or equivalently} \quad \mathbf{y} = P^{-1}\mathbf{x}$$

where P is a non-singular $n \times n$ matrix. It is easy to see $\mathbf{y} = [\mathbf{x}]_{\mathcal{B}}$, where $\mathcal{B}$ is the set of columns of P. Then

$$Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} = (P\mathbf{y})^\top A(P\mathbf{y}) = \mathbf{y}^\top(P^\top AP)\mathbf{y} = \mathbf{y}^\top D\mathbf{y} \tag{10.1}$$

which uses the fact that A is symmetric.

Example 10.1. Let

$$A = \begin{bmatrix} a & c \\ c & b \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

where A has eigenvalues λ_1 and λ_2. Then

$$Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} = a^2x_1^2 + b^2x_2 + 2cx_1x_2$$

By making the change of variable:

$$Q(\mathbf{x}) = Q'(\mathbf{y}) = \mathbf{y}^\top D\mathbf{y} = \lambda_1^2y_1^2 + \lambda_2^2y_2^2 \tag{10.2}$$

Remark 10.1. If we let $Q'(\mathbf{y}) = 1$, then $\lambda_1^2y_1^2 + \lambda_2^2y_2^2 = 1$ represents **an ellipse centred at the origin**.

Theorem 10.1. The Principal Axes Theorem: Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then there is an orthogonal change of variable, $\mathbf{x} = P\mathbf{y}$, that transforms the quadratic form $\mathbf{x}^\top A\mathbf{x}$ into a quadratic form $\mathbf{y}^\top D\mathbf{y}$ with no cross-product term. The columns of P are called the **principal axes** and $\mathbf{y}$ is the coordinate of $\mathbf{x}$ relative to the columns of P.

10.2 A Geometric View of Principal Axes

Suppose $Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} = k$, where A is 2×2 symmetric matrix and $k \in \mathbb{R}$. The set of all $\mathbf{x} \in \mathbb{R}^2$ that satisfy

$$\mathbf{x}^\top A\mathbf{x} = \mathbf{x}^\top A\mathbf{x} = k$$

It can be expressed as

$$a^2x_1^2 + bx_1x_2 + c^2x_2^2 + dx_1 + ex_2 + f = 0$$

which either corresponds to

1. an ellipse (or a circle):
$$a^2x_1^2 + bx_1x_2 + c^2x_2^2 + dx_1 + ex_2 + f = 0,\ ac > 0 \tag{10.3}$$
2. a hyperbola:
$$a^2x_1^2 + bx_1x_2 + c^2x_2^2 + dx_1 + ex_2 + f = 0,\ ac < 0 \tag{10.4}$$
3. two intersecting lines, if the equation (10.3) can be factorized to
$$(\alpha_1x_1 + \beta_1x_2 + \gamma_1)(\alpha_2x_1 + \beta_2x_2 + \gamma_2) = 0$$
4. a single point:
$$(x_1 - x_0)(x_2 - y_0) = 0$$

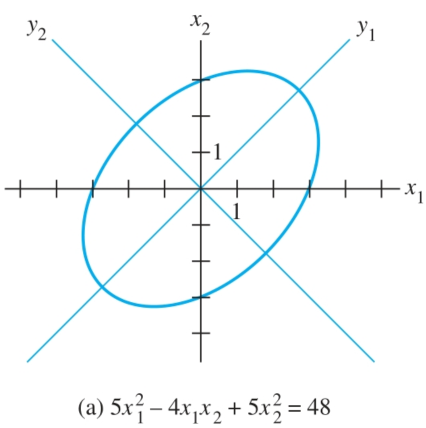
If A is a diagonal matrix, the graph is in *standard position*, which implies that the ellipse or the hyperbola is centred at the origin. Therefore, the equation (10.3) can be written as:

$$\frac{x_1}{a^2} + \frac{x_2}{b^2} = 1,\ a > 0,\ b > 0$$

The equation (10.4) can be written as:

$$\frac{x_1}{a^2} - \frac{x_2}{b^2} = 1,\ a > 0,\ b > 0$$

Find the *principal axes* (determined by the eigenvectors of A) amounts to finding a new coordinate system with respect to which the graph is in standard position (centred at the origin), as shown below:



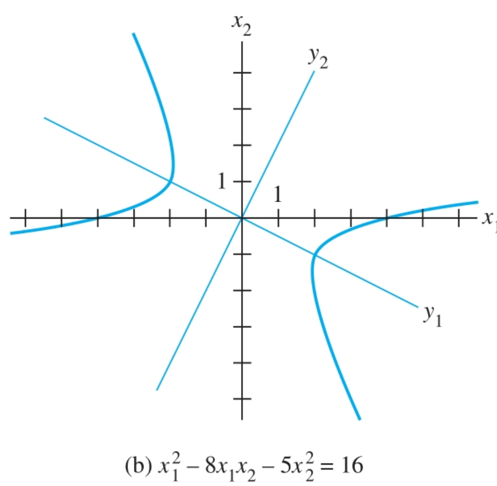


Figure 4: Finding the principal axes.

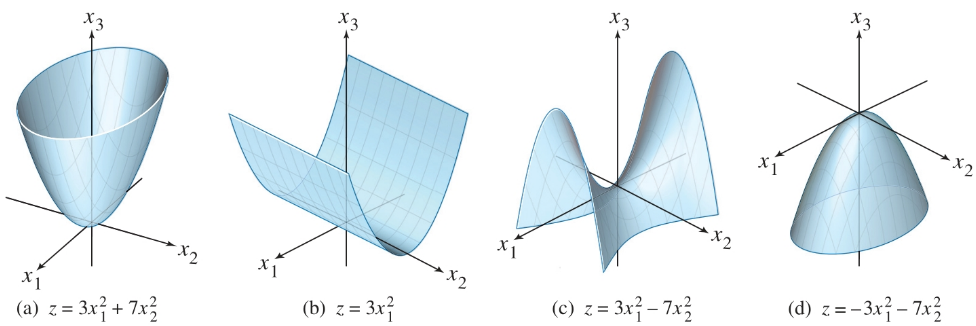


Figure 5: Graphs of quadratic forms.

10.3 Classifying Quadratic Forms

Definition 10.2. A quadratic from Q is

(a) **positive definite** if $\forall \mathbf{x} \neq \mathbf{0} : Q(\mathbf{x}) > 0$

(b) **negative definite** if $\forall \mathbf{x} \neq \mathbf{0} : Q(\mathbf{x}) < 0$

(c) **indefinite** if $Q(\mathbf{x})$ assumes both positive and negative values.

(d) **positive semi-definite** if $\forall \mathbf{x} : Q(\mathbf{x}) > 0$

(e) **negative semi-definite** if $\forall \mathbf{x} : Q(\mathbf{x}) < 0$

As shown in the figure 2.

Theorem 10.2. Quadratic Forms and Eigenvalues: Given a quadratic form $Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$, then Q is

a. positive definite if and only if the eigenvalues of A are all positive,

b. negative definite if and only if the eigenvalues of A are all negative, or

c. indefinite if and only A has both positive and negative eigenvalues.

Proof. By the equation (10.2),

$$Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} = \mathbf{y}^\top D\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2 \tag{10.5}$$

Since P is non-singular, there is a one-to-one relation between $\mathbf{x}$ and $\mathbf{y}$. For any nonzero $\mathbf{x}$, the right side of the equation above coincides with $Q(\mathbf{x})$ for $\mathbf{x} \neq \mathbf{0}$. Thereore, $Q(\mathbf{x})$ is obviously controlled by the signs of the eigenvalues of A, in the three ways described in the theorem. □

Remark 10.2. If A has a nonzero eigenvalue, say $\lambda_k = 0$, then $A\mathbf{x} = 0$ has a non-trivial solution, implying $\exists \mathbf{x} \neq \mathbf{0} : Q(\mathbf{x}) = 0$.

11 TODO: A preview of Constrained Optimization

11.1 Subject to a Unit Vector

In some applications, we often need to find the maximum or minimum value of a quadratic form $Q(\mathbf{x})$ for $\mathbf{x}$ in some specified set. For example,

$$c = \operatorname*{argmin}_{\|\mathbf{x}\|=1} Q(\mathbf{x})$$

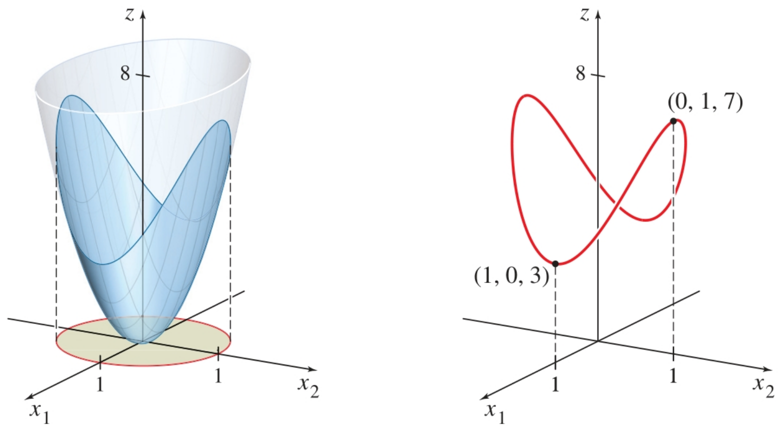


Figure 6: $z = 3x_1^2 + 7x_2^2$ constrained on $x_1^2 + x_2^2 = 1$

Theorem 11.1. Given a quadratic form $Q(\mathbf{x})$, and let $m = \underset{\|\mathbf{x}\|=1}{\operatorname{argmin}} Q(\mathbf{x})$ and $M = \underset{\|\mathbf{x}\|=1}{\operatorname{argmax}} Q(\mathbf{x})$. then

1. M is the greatest eigenvalue λ_1 of A
2. m is the least eigenvalue λ_n of A.

The value of $\mathbf{x}^\top A\mathbf{x}$ is

1. M when $\mathbf{x}$ is a unit eigenvector $\mathbf{u}_1$ corresponding to λ_1
2. m when $\mathbf{x}$ is a unit eigenvector $\mathbf{u}_m$ corresponding to λ_n

Proof. By the theorem 9.10, A can be orthogonally diagonalized as PDP^{-1}, where either P or P^{-1} is an orthogonal matrix, thus preserving the length $\mathbf{x}$. By equation (10.2)

$$Q(\mathbf{x}) = Q'(\mathbf{y}) = \sum_{i=1}^{n} \lambda_i y_i^2$$

where λ's are arranged in descending order. The following inequality holds:

$$Q'(\mathbf{y}) \leq \lambda_1 \sum_{i=1}^{n} y_i^2 = \lambda_1 \mathbf{y}^\top \mathbf{y}$$

where λ_1 is the largest eigenvalue of A. Let $\mathbf{y}$ be $\mathbf{e}_1$, a vector with the first entry being 1 and the other being 0. Then,

$$\lambda_1 \mathbf{y}^\top \mathbf{y} = \mathbf{e}_1^\top D\mathbf{e}_1$$

illustrates that $Q'(\mathbf{y})$ reaches its maximum value when $\mathbf{y} = \mathbf{e}_1$, implying that $Q(\mathbf{x})$ attains its maximum value when $\mathbf{x} = P\mathbf{e}_1 = \mathbf{u}_1$. A similar method can be applied to prove its minimum value. $\square$

Theorem 11.2. Given a quadratic form $Q(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$, let λ_1 be the largest eigenvalue of A, and $\mathbf{u}_1$ be the eigenvector corresponding to λ_1. then the maximum value of Q subject to the following constrains:

$$\mathbf{x}^\top \mathbf{x} = 1, \quad \mathbf{x}^\top \mathbf{u}_1 = 0$$

is the second greatest eigenvalue λ_2, and this maximum is attained when $\mathbf{x}$ is an eigenvector $\mathbf{u}_2$ corresponding to λ_2.

Remark 11.1. Suppose that A is orthogonally diagonalized as PDP^{-1} with its eigenvalues arranged, in descending order, on the main diagonal of D. If there are more constrains on Q:

$$\mathbf{x}^\top \mathbf{x} = 1, \quad \mathbf{x}^\top \mathbf{u}_1 = 0, \quad \cdots, \mathbf{x}^\top \mathbf{u}_{k-1} = 0$$

then the maximum of Q is attained at $\mathbf{x} = \mathbf{u}_k$ where $\mathbf{u}_k$ is the eigenvector corresponding to the k^{th} greatest eigenvalue.

12 TODO: Singular Value Decomposition

Unfortunately, as we know, not all matrices can be factored as $A = PDP^{-1}$ with D diagonal. However, a factorization $A = QDP^{-1}$ is possible for *any* $m \times n$ matrix A! A special factorization of this type, called the **singular value decomposition**, is **the most useful matrix decomposition in the universe.**😉

If $A\mathbf{x} = \lambda\mathbf{x}$ and $\|\mathbf{x}\| = 1$, then

$$\|A\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\| = |\lambda|$$

If λ_1 is the eigenvalue with the greatest magnitude, then a corresponding unit eigenvector $\mathbf{v}_1$ identifies a direction in which the stretching effect of A is greatest.

Example 12.1. If the linear transformation $\mathbf{x} \to A\mathbf{x}$ maps the unit sphere $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$ in $\mathbb{R}^3$ onto an ellipse in $\mathbb{R}^2$. Find a unit vector $\mathbf{x}$ at which the length $\|A\mathbf{x}\|$ is maximized, and compute this maximum length.

Solution. The quantity $\|A\mathbf{x}\|^2$ is maximized at the same $\mathbf{x}$ that maximizes $\|A\mathbf{x}\|$,

$$\|A\mathbf{x}\|^2 = (A\mathbf{x})^\top (A\mathbf{x}) = \mathbf{x}^\top (A^\top A)\mathbf{x}$$

Since $A^\top A$ is symmetric, so the problem is reduced into maximizing the quadratic form $\mathbf{x}^\top (A^\top A)\mathbf{x}$ subject to the constraint $\|\mathbf{x}\| = 1$ as discussed in theorem 11.1. Hence, the maximum value is the greatest eigenvalue λ_1 of $A^\top A$, and the maximum value is attained at a unit eigenvector of $A^\top A$ corresponding to λ_1.

The example above suggests that the effect of A on the unit sphere in $\mathbb{R}^3$ is related to the quadratic form $x^\top (A^\top A)\mathbf{x}$.

Let $A \in \mathbb{R}^{m\times n}$. Then $A^\top A$ can be orthogonally diagonalized. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $A^\top A$. Then,

$$\|A\mathbf{v}_i\| = (A\mathbf{v}_i)^\top A\mathbf{v}_i = \mathbf{v}_i^\top (\lambda_i \mathbf{v}_i) = \lambda_i \geq 0 \tag{12.1}$$

Note that $\|\mathbf{v}_i\| = 1$. So **the eigenvalues of $A^\top A$ are all non-negative, implying that $A^\top A$ is a semi-positive definite matrix.**

Definition 12.1. The singular values of A are the square roots of the eigenvalues of $A^\top A$, denoted by $\sigma_1, \cdots, \sigma_n$, and they are arranged in decreasing order. By equation (12.1), the **singular values of A are the lengths of the vectors $A\mathbf{v}_1, \cdots, A\mathbf{v}_n$.**

Remark 12.1. The first two singular values of A are the lengths of the major and minor semi-axes of the ellipse as shown figure 7.

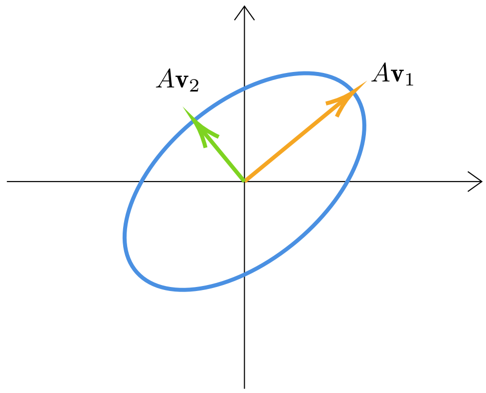


Figure 7: $A\mathbf{v}_1$ is the major semi-axis and $A\mathbf{v}_2$ is the minor semi-axis of the ellipse.

Theorem 12.1. Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of $A^\top A$, arranged so that the corresponding eigenvalues of $A^\top A$ satisfy $\lambda_1 \geq \cdots \geq \lambda_n$, and suppose that A has r nonzero singular values. Then $\{A\mathbf{v}_1, \cdots, A\mathbf{v}_r\}$ is an orthogonal basis for $\mathrm{Col}(A)$, and $\mathrm{rank}(A) = r$.

Proof. Given two vectors $A\mathbf{v}_j, A\mathbf{v}_i$ where $i \neq j$,

$$(A\mathbf{v}_j)^\top A\mathbf{v}_i = \mathbf{v}_j^\top A^\top A\mathbf{v}_i = \lambda_i \mathbf{v}_j^\top \mathbf{v}_i = 0$$

Thus, $\{A\mathbf{v}_1, \cdots, A\mathbf{v}_n\}$ is an orthogonal set. Furthermore, since the lengths of the vector $\{A\mathbf{v}_1, \cdots, A\mathbf{v}_n$ are the singular values of A, and since there are r non-zero singular values, $A\mathbf{v}_i \neq \mathbf{0}$ if and only if $1 \leq i \leq r$. So, $\{A\mathbf{v}_1, \cdots, A\mathbf{v}_r$ are linearly independent vectors, and they are in $\mathrm{Col}(A)$. $\forall \mathbf{y} \in \mathrm{Col}(A)$, say $\mathbf{y} = A\mathbf{x}$, we can write

$$\mathbf{x} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$$

, and

$$\begin{aligned}
\mathbf{y} &= A\mathbf{x} \\
&= c_1 A\mathbf{v}_1 + \cdots + c_r A\mathbf{v}_r + c_{r+1} A\mathbf{v}_{r+1} + \cdots + c_n A\mathbf{v}_n \\
&= c_1 A\mathbf{v}_1 + \cdots + c_r A\mathbf{v}_r + 0 + 0 + \cdots + 0
\end{aligned}$$

Thus $\mathbf{y}$ is in $\mathrm{Span}\{A\mathbf{v}_1, \cdots, A\mathbf{v}_r\}$, which shows that $\{A\mathbf{v}_1, \cdots, A\mathbf{v}_r\}$ is an (orthogonal) basis for $\mathrm{Col}(A)$. Hence $\mathrm{rank}(A) = \dim\Big(\mathrm{Col}(A)\Big) = r$. $\square$

The decomposition of A involves an $m \times n$ “diagonal” matrix Λ of the form

$$\Lambda = \begin{bmatrix} D & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \tag{12.2}$$

where D is an $r \times r$ diagonal matrix for some r not exceeding the smaller of m and n.

Theorem 12.2. The Singular Value Decomposition or (SVD): Let A be an $m \times n$ matrix with rank r. then there exists an $m \times n$ matrix Λ as in equation (12.2) for which the diagonal entries in D are the first singular values of A, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$, and there exist an $m \times n$ orthogonal matrix U and an $n \times n$ orthogonal matrix V such that

$$A = U\Lambda V^\top$$

Proof. Let λ_i and $\mathbf{v}_i$ be as in theorem 12.1, so that $\{A\mathbf{v}_1, \cdots, A\mathbf{v}_r\}$ is an orthogonal basis for $\mathrm{Col}(A)$. Normalize each $A\mathbf{v}_i$ to obtain an orthonormal basis $\mathcal{U} = \{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$, where

$$\mathbf{u}_i = \frac{A\mathbf{v}_i}{\|A\mathbf{v}_i\|} = \frac{A\mathbf{v}_i}{\sigma_i} \tag{12.3}$$

and

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i \tag{12.4}$$

Now extend $\mathcal{U}$ to an orthonormal basis $\{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$ of $\mathbb{R}^m$, and let

$$U = \begin{bmatrix}\mathbf{u_1} & \cdots & \mathbf{u}_m\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_n\end{bmatrix} \tag{12.5}$$

By construction, U and V are orthogonal matrices. Also,

$$AV = \begin{bmatrix}A\mathbf{v}_1 & \cdots & A\mathbf{v}_r & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\sigma_1\mathbf{u}_1 & \cdots & \sigma_r\mathbf{u}_r & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix} \tag{12.6}$$

Let D be the diagonal matrix with diagonal entries $\sigma_1, \cdots, \sigma_r$, and let Λ be as in theorem 12.1 above. Then

$$U\Lambda = \begin{bmatrix}U_1 & U_2\end{bmatrix}\begin{bmatrix}D & \mathbf{O} \\ \mathbf{O} & \mathbf{O}\end{bmatrix} = \begin{bmatrix}U_1 D & \mathbf{O}\end{bmatrix} = AV \tag{12.7}$$

where $U_1 = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_r\end{bmatrix}$ and $U_2 = \begin{bmatrix}\mathbf{u}_{r+1} & \cdots & \mathbf{u}_m\end{bmatrix}$. Since V is an orthogonal matrix,

$$U\Lambda V^\top = AVV^\top = A$$

□

Remark 12.2. The columns of U are called **left singular vectors** of A, and the columns of V are called **right singular vectors** of A.

12.1 Bases for Fundamental Subspaces

Given an SVD decomposition for a $m \times n$ matrix A, by observing its left singular vectors, we can find that $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ is an orthonormal basis for $\mathrm{Col}(A)$ by theorem 12.1, and $\{\mathbf{u}_{r+1}, \cdots, \mathbf{u}_n\}$ is an orthonormal basis for $\mathrm{Nul}(A^\top)$, since for any $r < i \leq n$, $\mathbf{u}_i$ is orthogonal to $\mathrm{Col}(A) = \mathrm{Span}\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$, that is, $\mathrm{Span}\{\mathbf{u}_{r+1}, \cdots, \mathbf{u}_m\} = \mathrm{Col}(A)^\perp$.

Since $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ forms a basis for $\mathrm{Col}(A)$, $\dim(A) = r$, implying $\dim\Big(\mathrm{Nul}(A)\Big) = n - r$. For any $i > r$, since $A\mathbf{v}_i = \mathbf{0}$ and $\dim\Big(\mathrm{Nul}(A)\Big) = n - r$, $\mathrm{Span}\{\mathbf{v}_{r+1}, \cdots, \mathbf{v}_n\} = \mathrm{Nul}(A)$. Note that $\mathrm{Nul}(A)^\perp = \mathrm{Row}(A)$. Hence, $\{\mathbf{v}_1, \cdots, v_r\}$ is an orthonormal basis for $\mathrm{Row}(A)$. Observing that

$$AV = \begin{bmatrix}A\mathbf{v}_1 & \cdots & A\mathbf{v}_r & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\sigma_1\mathbf{u}_1 & \cdots & \sigma_r\mathbf{u}_r & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix}$$

for which the non-zero vectors of AV is an orthogonal basis for $\mathrm{Col}(A)$. In other words, the matrix A transforms a collection of basis vectors of $\mathrm{Col}(A)$ and $\mathrm{Nul}(A^\top)$ into a collection of basis of $\mathrm{Row}(A)$ and $\mathrm{Nul}(A)$.

Let $V = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_r\end{bmatrix}$, $\begin{bmatrix}\mathbf{v}_{r+1} & \cdots & \mathbf{v}_n\end{bmatrix}$. And let $U_1 = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_r\end{bmatrix}$, $\begin{bmatrix}\mathbf{u}_{r+1} & \cdots & \mathbf{u}_m\end{bmatrix}$, they have the relationship shown as figure 8,

13 Vector Calculus

13.1 Gradient

$$\mathrm{Row}(A) = \mathrm{Span}(V_1) \xrightarrow{A} \mathrm{Span}(U_1) = \mathrm{Col}(A)$$

$$\mathrm{Nul}(A) = \mathrm{Span}(V_2) \xrightarrow{A} \mathrm{Span}(U_2) = \mathrm{Nul}\left(A^\top\right)$$

Figure 8: The effect of A on V.

Definition 13.1. Gradient: Let $f : \mathbb{R}^n \to \mathbb{R}$. The gradient of the function f with respect to $\mathbf{x}$ is a vector of n partial derivatives:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \partial_{x_1} f(\mathbf{x}) & \partial_{x_2} f(\mathbf{x}) & \cdots & \partial_{x_n} f(\mathbf{x}) \end{bmatrix}^\top \tag{13.1}$$

$\nabla_{\mathbf{x}} f(\mathbf{x})$ is typically replaced by $\nabla f(\mathbf{x})$.

The following rules come in handy for differentiating multivariate function:

1. $\forall A \in \mathbb{R}^{n \times p}$: $\nabla_{\mathbf{x}} A\mathbf{x} = A^\top$
2. $\forall A \in \mathbb{R}^{p \times p}$: $\nabla_{\mathbf{x}} \mathbf{x}^\top A \mathbf{x} = (A + A^\top)\mathbf{x}$
3. $\nabla_{\mathbf{x}} \|\mathbf{x}\|^2 = \nabla_{\mathbf{x}} \mathbf{x}^\top \mathbf{x} = 2\mathbf{x}$

Theorem 13.1. Chain Rule: Suppose $y = f(\mathbf{u})$ has variables $u_1, u_2, \cdots, u_m$. where each $u_i = g_i(\mathbf{x})$ has variables $x_1, x_2, \cdots, x_n$, i.e., $\mathbf{u} = g(\mathbf{x})$. Then

$$\frac{\partial y}{\partial x_i} = \frac{\partial y}{\partial u_1}\frac{\partial u_1}{\partial x_i} + \frac{\partial y}{\partial u_2}\frac{\partial u_2}{\partial x_i} + \cdots + \frac{\partial y}{\partial u_m}\frac{\partial u_m}{\partial x_i} = A\nabla_{\mathbf{u}} y. \tag{13.2}$$

where $A \in \mathbb{R}^{n \times m}$ contains the derivative of vector $\mathbf{u}$ with respect to vector $\mathbf{x}$.

Example 13.1. Let X be an $n \times p$ matrix, find a vector $\hat{\boldsymbol{\beta}}$ such that

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{y} - X\mathbf{b}\|^2$$

Solution. Let $f(\mathbf{b}) = \|\mathbf{y} - X\mathbf{b}\|^2 = (\mathbf{y} - X\mathbf{b})^\top(\mathbf{y} - X\mathbf{b})$. Expanding $(\mathbf{y} - X\mathbf{b})^\top(\mathbf{y} - X\mathbf{b})$.

$$f(\mathbf{b}) = \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top X\mathbf{b} - \mathbf{b}^\top X^\top \mathbf{y} + \mathbf{b}^\top X^\top X \mathbf{b}$$

It is easy to see the above equation has a minimum value. Let its gradient be $\mathbf{0}$:

$$\nabla f(\mathbf{b}) = -X^\top \mathbf{y} - X^\top \mathbf{y} + 2X^\top X\mathbf{b} = \mathbf{0}$$

If $X^\top X$ is non-singular, we can get $\mathbf{b} = (X^\top X)^{-1} X^\top \mathbf{y}$.

13.2 Jacobin Matrix

Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable function on region $D \subseteq \mathbb{R}^m$. That is, $\forall \mathbf{x} \in D$,

$$F(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & \cdots & f_n(\mathbf{x}) \end{bmatrix}^\top$$

for which each f_i is an $\mathbb{R}^n \to \mathbb{R}$ function. However, since the function F can be arbitrarily complected, a good approach is to find a linear function that approximate F around a point $\mathbf{p} \in \mathbb{R}^n$. Suppose we can find such a function, say $T(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$. It must satisfy the following conditions:

1. $F(\mathbf{p}) = T(\mathbf{p})$
2. $\lim_{\mathbf{x}\to\mathbf{p}} F(\mathbf{p}) - T(\mathbf{x}) = 0$

By the first condition, $T(\mathbf{p}) = A\mathbf{p} + \mathbf{b}$, we have

$$\mathbf{b} = F(\mathbf{p}) - A\mathbf{p} \tag{13.3}$$

Substitute equation above to $T(\mathbf{x})$

$$T(\mathbf{x}) = A\mathbf{x} + F(\mathbf{p}) - A\mathbf{p} = F(\mathbf{p}) + A(\mathbf{x} - \mathbf{p}) \tag{13.4}$$

Then, the condtion 2 can be written as

$$\lim_{\mathbf{x}\to\mathbf{p}} F(\mathbf{x}) - F(\mathbf{p}) + A(\mathbf{x} - \mathbf{p}) = 0 \tag{13.5}$$

We can handle a simpe case with it: let $\mathbf{x}$ approaches to $\mathbf{p}$ along a standard coordinate axis. Let $\mathbf{e}_i$ be a vector, where the i^{th} entry is one and all other entries are zeros, and $\mathbf{x} = \mathbf{p} + h\mathbf{e}_j$. Then

$$\lim_{h\to 0} F(\mathbf{p} + h\mathbf{e}_j) - F(\mathbf{p}) + A(h\mathbf{e}_j) = 0 \tag{13.6}$$

where $h \neq 0$. The equation above is equiavalent to

$$\lim_{h\to 0} \frac{F(\mathbf{p} + h\mathbf{e}_j) - F(\mathbf{p}) + A(h\mathbf{e}_j)}{h} = \lim_{h\to 0} \frac{F(\mathbf{p} + h\mathbf{e}_j) - F(\mathbf{p}) + hA(h\mathbf{e}_j)}{h} = 0 \tag{13.7}$$

We can get

$$\lim_{h\to 0} \frac{F(\mathbf{p} + h\mathbf{e}_j) - F(\mathbf{p})}{h} = A\mathbf{e}_j = \begin{bmatrix} \frac{\partial f_1}{\partial x_j}(\mathbf{p}) & \frac{\partial f_2}{\partial x_j}(\mathbf{p}) & \cdots & \frac{\partial f_m}{\partial x_j}(\mathbf{p}) \end{bmatrix}^\top \tag{13.8}$$

Hence,

$$A = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{p}) & \frac{\partial f_2}{\partial x_1}(\mathbf{p}) & \cdots & \frac{\partial f_n}{\partial x_1}(\mathbf{p}) \\ \frac{\partial f_1}{\partial x_2}(\mathbf{p}) & \frac{\partial f_2}{\partial x_n}(\mathbf{p}) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{p}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_m}(\mathbf{p}) & \frac{\partial f_n}{\partial x_m}(\mathbf{p}) & \cdots & \frac{\partial f_n}{\partial x_m}(\mathbf{p}) \end{bmatrix} \tag{13.9}$$

Note that the matrix A discussed in theorem 13.1 has the similar form as the above matrix.

Definition 13.2. Jacobin Matrix: Suppose $\mathbf{y} = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$ is a continuous function with continuous partial derivatives, where each $y_i = f_i(\mathbf{x})$. Its Jacobin matrix is defined as below:

$$J_f = \nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_n}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_m} & \frac{\partial f_n}{\partial x_m} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \tag{13.10}$$

Theorem 13.2. Suppose a function $f : \mathbb{R}^n \to \mathbb{R}^n$ as discussed in definition 13.2 is invertible, then

$$\det(J_{f^{-1}}) = \det(J_f)^{-1} \tag{13.11}$$

13.3 Multivariate Taylor's Theorem

We've learned Taylor series for a function $y = f(x)$ of a single variable. For an $n+1$-times differentiable function $f : \mathbb{R} \to \mathbb{R}$, we have

$$f(x) = f(c) + f'(c)(x-c) + \frac{f''(c)(x-c)^2}{2!} + \cdots + \frac{f^{(n)}(c)(x-c)^n}{n!} + R_n(x,c) \tag{13.12}$$

where $R_n(x,c)$ is called the **remained term**:

$$R_n(x,c) = \frac{f^{(n+1)}(z)(x-c)^n}{n!} \tag{13.13}$$

for which z is a real number between x and c. There is a very similar formula for functions of several variables. Before go further, let us define some notations. For a function $f : \mathbb{R}^n \to \mathbb{R}$ and two vectors: $\mathbf{x}_0, \mathbf{h} \in \mathbb{R}^n$:

$$\begin{aligned}
D_f(\mathbf{x}_0, \mathbf{h}) &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}_0) h_i \\
D_f^2(\mathbf{x}_0, \mathbf{h}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0) h_i h_j \\
D_f^3(\mathbf{x}_0, \mathbf{h}) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(\mathbf{x}_0) h_i h_j h_k
\end{aligned}$$

and so on. Note that $D_f(\mathbf{x}_0, \mathbf{h}) = \nabla f(\mathbf{x}_0)^\top \mathbf{h}$.

Theorem 13.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an $n+1$-times continuously differentiable function at the point $\mathbf{v_0} \in \mathbb{R}^n$. Then,

$$f(\mathbf{x}) = f(\mathbf{v}_0) + \sum_{k=1}^{n} \frac{1}{k!} D_f^k(\mathbf{v}_0, \mathbf{x} - \mathbf{v}_0) + \frac{1}{(n+1)!} D_f^{n+1}(\mathbf{z}, \mathbf{x} - \mathbf{v}_0)$$

where $\mathbf{z}$ is some point on the segment from $\mathbf{x}$ to $\mathbf{v}_0$.

Example 13.2. Write out the Taylor expansion through terms of degree 2 for $f : \mathbb{R}^2 \to \mathbb{R}^2$. Let $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$

$$\begin{aligned}
f(\mathbf{x}) = f(\mathbf{v}_0) &+ \left(\frac{\partial f}{\partial x_1}(\mathbf{v}_0)(x_1 - v_1) + \frac{\partial f}{\partial x_2}(\mathbf{v}_0)(x_2 - v_2) \right) + \\
&\frac{1}{2}\left(\frac{\partial^2 f}{\partial x_1^2}(\mathbf{v}_0)(x_1 - v_1)^2 + \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{v}_0)(x_1 - v_1)(x_2 - v_2) + \frac{\partial^2 f}{\partial x_2^2}(\mathbf{v}_0)(x_2 - v_2)^2 + \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{v}_0)(x_2 - v_2)(x_1 - v_1) \right) \\
&+ \cdots
\end{aligned}$$

Note the term of degree 1 can be written as $\nabla f(\mathbf{v_0})^\top(\mathbf{x} - \mathbf{v}_0)$, and the term of degree 2 can be written in a quadratic form:

$$\frac{1}{2}(\mathbf{x} - \mathbf{v}_0)^\top \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{v}_0) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{v}_0) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{v}_0) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{v}_0) \end{bmatrix} (\mathbf{x} - \mathbf{v}_0) \tag{13.14}$$

The matrix in equation (13.14) is called **Hessian Matrix**, denoted by $\mathbf{H}_f(\mathbf{v}_0)$.

13.4 TODO: Hessian Matrix

Definition 13.3. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ has continuous second-order derivatives. Then the Hessian Matrix is a square $n \times n$ matrix, usually defined and arranged as

$$H_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{13.15}$$

13.5 TODO: Put them together

14 Probability and Statistics

In this section, an uppercase letter (e.g., X) represents a random variable, while a bold upper letter (e.g., $\mathbf{X}$) represents a random vector, random matrix, or real matrix.

14.1 Expectation and Variance for Random Matrix

Definition 14.1. The expectation of a random vector $\mathbf{X} \in \mathbb{R}^p$ is a p-dimensional vector defined as:

$$\mathbb{E}(\mathbf{X}) = \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_p) \end{bmatrix} = \boldsymbol{\mu}_{\mathbf{X}}$$

Definition 14.2. Covariance Matrix: A $p \times p$ matrix $\mathbf{\Sigma}$ defined as

$$\mathbf{\Sigma} = \mathrm{Var}(\mathbf{X}) = \mathbb{E}((\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top)$$

is called the covariance matrix of $\mathbf{X}$. We can expand the outer product:

$$\mathrm{Var}(\mathbf{X}) = \begin{bmatrix} \mathbb{E}\Big((X_1 - \mu_1)^2\Big) & \mathbb{E}\Big((X_1 - \mu_1)(X_2 - \mu_2)\Big) & \cdots & \mathbb{E}\Big((X_1 - \mu_1)(X_p - \mu_p)\Big) \\ \vdots & \ddots & \cdots & \vdots \\ \mathbb{E}\Big((X_p - \mu_p)(X_1 - \mu_1)\Big) & \mathbb{E}\Big((X_p - \mu_p)(X_2 - \mu_2)\Big) & \cdots & \mathbb{E}\Big((X_p - \mu_p)^2\Big) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

where σ_{ij} stands for $\mathrm{Cov}(X_i, X_j)$. It is easy to see that **$\mathrm{Var}(\mathbf{X})$ is symmetric** due to the fact that $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$.

Remark 14.1. If $\mathbf{X}$ and $\mathbf{Y}$ are two random vectors with different joint probability distributions, then $\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$ is **NOT symmetric**, therefore $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) \neq \mathrm{Cov}(\mathbf{Y}, \mathbf{X})$.

Theorem 14.1. $\mathrm{Var}(\mathbf{X})$ has the following equivalent representation:

$$\mathrm{Var}(\mathbf{X}) = \mathbb{E}(\mathbf{X}\mathbf{X}^\top) - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top \tag{14.1}$$

due to the fact that $\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)$.

Theorem 14.2. The following rules come in handy for calculating expectation and variance:

1. $\mathbb{E}(\mathbf{X}+\mathbf{C}) = \mathbb{E}(\mathbf{X}) + \mathbf{C}$, where $\mathbf{X}$ is an $n \times p$ random matrix and $\mathbf{C} \in \mathbb{R}^{n\times p}$.
2. $\mathbb{E}(\mathbf{A}\mathbf{X}+\mathbf{C}) = \mathbf{A}\mathbb{E}(\mathbf{X}) + \mathbf{C}$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{C} \in \mathbb{R}^{m\times p}$.
3. $\mathbb{E}(\mathbf{Q}\mathbf{X}\mathbf{P}) = \mathbf{Q}\mathbb{E}(\mathbf{X})\mathbf{P}$, where $\mathbf{Q}, \mathbf{P}$ are properly defined real matrices.
4. $\mathbb{E}(\mathbf{Q}\mathbf{X}^\top\mathbf{P}) = \mathbf{Q}\mathbb{E}(\mathbf{X})^\top\mathbf{P}$, where $\mathbf{Q}, \mathbf{P}$ are properly defined.
5. $\mathbb{E}(\mathbf{Q}\mathbf{X}\mathbf{P}+\mathbf{b})^\top = \mathbb{E}\Big((\mathbf{Q}\mathbf{X}\mathbf{P}+\mathbf{b})^\top\Big)$
6. $\mathrm{Var}(\mathbf{A}\mathbf{X}+\mathbf{b}) = \mathbf{A}\mathrm{Var}(\mathbf{X})\mathbf{A}^\top$

   Proof.

$$\mathbb{E}(\mathbf{A}\mathbf{X}+\mathbf{b})\mathbb{E}(\mathbf{A}\mathbf{X}+\mathbf{b})^\top = \mathbf{A}\boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\mathbf{A}^\top + \mathbf{A}\boldsymbol{\mu}_\mathbf{X}\mathbf{b}^\top + \mathbf{b}\boldsymbol{\mu}_\mathbf{X}^\top\mathbf{A}^\top + \mathbf{b}\mathbf{b}^\top \tag{14.2}$$

$$\mathbb{E}\Big((\mathbf{A}\mathbf{X}+\mathbf{b})(\mathbf{A}\mathbf{X}+\mathbf{b})^\top\Big) = \mathbf{A}\mathbb{E}(\mathbf{X}\mathbf{X}^\top)\mathbf{A}^\top + \mathbf{A}\boldsymbol{\mu}_\mathbf{X}\mathbf{b}^\top + \mathbf{b}\boldsymbol{\mu}_\mathbf{X}^\top\mathbf{A}^\top + \mathbf{b}\mathbf{b}^\top \tag{14.3}$$

   By subtracting equation (14.3) by equation (14.2),

$$\begin{aligned}\mathrm{Var}(\mathbf{A}\mathbf{X}+\mathbf{b}) &= \mathbf{A}\mathbb{E}(\mathbf{X}\mathbf{X}^\top)\mathbf{A}^\top - \mathbf{A}\boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\mathbf{A}^\top \\ &= \mathbf{A}\Big(\mathbb{E}(\mathbf{X}\mathbf{X}^\top) - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\Big)\mathbf{A}^\top \\ &= \mathbf{A}\mathrm{Var}(\mathbf{X})\mathbf{A}^\top\end{aligned}$$

   □

7. $\mathrm{Cov}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) = \mathbf{A}\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^\top$
8. $\mathbb{E}(\mathbf{X}^\top\mathbf{A}\mathbf{X}) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}) + \boldsymbol{\mu}_\mathbf{X}^\top\mathbf{A}\boldsymbol{\mu}_\mathbf{X}$

   Proof. The proof uses the properties discussed in theorem 8.1.

$$\begin{aligned}\mathbb{E}(\mathbf{X}^\top\mathbf{A}\mathbf{X}) &= \mathbb{E}\Big(\mathrm{tr}(\mathbf{X}^\top\mathbf{A}\mathbf{X})\Big) \quad \text{Since}\mathbf{X}^\top\mathbf{A}\mathbf{X} \text{ is a scalar.} \\ &= \mathbb{E}\Big(\mathrm{tr}(\mathbf{A}\mathbf{X}\mathbf{X}^\top)\Big) \\ &= \mathrm{tr}\Big(\mathbb{E}(\mathbf{A}\mathbf{X}\mathbf{X}^\top)\Big) = \mathrm{tr}\Big(A\mathbb{E}(\mathbf{X}\mathbf{X}^\top)\Big) \\ &= \mathrm{tr}\Big(\mathbf{A}(\boldsymbol{\Sigma}_\mathbf{X} + \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top))\Big) \\ &= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}) + \mathrm{tr}(\mathbf{A}\boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}) + \mathrm{tr}(\boldsymbol{\mu}_\mathbf{X}\mathbf{A}\boldsymbol{\mu}_\mathbf{X}^\top) \\ &= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}) + \boldsymbol{\mu}_\mathbf{X}^\top\mathbf{A}\boldsymbol{\mu}_\mathbf{X}\end{aligned}$$

   □

The property 8 is useful when calculating expectation involving a quadratic from.

Example 14.1. Suppose $Y = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix}^\top$ is a random vector where Y_i's are i.i.d. distributed with mean μ and variance σ^2. Then $\mathbb{E}(\mathbf{Y}) = \mu\mathbb{1}$ and $\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}$. $\displaystyle\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ can be expressed in a quadratic form: $\mathbf{Y}^\top(\mathbf{I} - \mathbf{H}_0)\mathbf{Y}$, where $\mathbf{H}_0$ is the projection matrix onto vector $\mathbb{1}$. Note that $\mathbf{H}_0$ is full of $\dfrac{1}{n}$'s.

$$
\begin{aligned}
E\Big(\mathbf{Y}^\top(\mathbf{I} - \mathbf{H}_0)\mathbf{Y}\Big) &= \mathrm{tr}\Big((\mathbf{I} - \mathbf{H}_0)\sigma^2\mathbf{I}\Big) + \boldsymbol{\mu}_\mathbf{Y}^\top(\mathbf{I} - \mathbf{H}_0)\boldsymbol{\mu}_\mathbf{Y} \\
&= \sigma^2(1 - \frac{1}{n})n + \mu^2\mathbb{1}^\top(\mathbf{I} - (\mathbb{1}^\top\mathbb{1})^{-1}\mathbb{1}\mathbb{1}^\top)\mathbb{1} \\
&= \frac{\sigma^2}{n-1} + \mu^2(\mathbb{1}^\top - \mathbb{1}^\top)\mathbb{1} \\
&= \frac{\sigma^2}{n-1}
\end{aligned}
$$

We can see that $\dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$ is an unbiased estimator.

Theorem 14.3. The covariance matrix $\boldsymbol{\Sigma}$ of a random vector $\mathbf{X} \in \mathbb{R}^n$ is **positive semi-definite** as discussed in definition 10.2.

Proof. Let $Y = \mathbf{b}^\top(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})$, where $\mathbf{b} \in \mathbb{R}^n$, then

$$
\begin{aligned}
\mathbb{E}(Y^2) &= \mathbb{E}(YY^\top) \\
&= \mathbb{E}\Big(\mathbf{b}^\top(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})^\top\mathbf{b}\Big) \\
&= \mathbf{b}^\top\boldsymbol{\Sigma}\mathbf{b} \geq 0
\end{aligned}
$$

□

This theorem illustrates that the eigenvalues of $\boldsymbol{\Sigma}$ are non-negative, and, therefore, $\det() \geq 0$. $\boldsymbol{\Sigma}$ is positive definite if and only if all of its eigenvalues are positive by theorem 10.2, implying that $\det() > 0$.

14.2 Transformations for Random Vectors

Theorem 14.4. Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector, with joint p.d.f. $f_\mathbf{X}(\mathbf{x})$. Let $G : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous and invertible function with continuous partial derivatives. If we let $\mathbf{Y} = G(\mathbf{X})$, then $\mathbf{Y}$ is also a random vector with joint p.d.f.

$$
f_\mathbf{Y}(\mathbf{y}) = f_\mathbf{X}\Big(G^{-1}(\mathbf{Y})\Big)|\det(\mathbf{J}_{G^{-1}})| \tag{14.4}
$$

where $\mathbf{J}_{G^{-1}}$ is a Jacobin matrix defined as definition 13.2.

Example 14.2. Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector, with joint p.d.f. $f_{\mathbf{X}}(\mathbf{x})$. Let $\mathbf{Y} = G(\mathbf{X}) = \mathbf{AX} + \mathbf{b}$, where $\mathbf{A}$ is an $n \times n$ non-singular real matrix. Find the joint p.d.f. of $\mathbf{Y}$.

Solution. Obviously, the linear transformation $\mathbf{AX}+\mathbf{b}$ is invertible, since $\mathbf{A}^{-1}$ exists. Therefor, $\mathbf{X} = \mathbf{A}^{-1}(\mathbf{Y}-\mathbf{b})$ with Jacobin matrix:

$$\mathbf{J}_{G^{-1}} = \nabla G(\mathbf{Y})^{-1} = (\mathbf{A}^{-1})^\top \tag{14.5}$$

Thus,

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\Big(G^{-1}(\mathbf{Y})\Big)|\det(A^{-1})| \tag{14.6}$$

The result can also be written as $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\Big(G^{-1}(\mathbf{Y})\Big)|\det(A)|^{-1}$.

14.3 Multivariate Gaussian Distribution

We know that the linear combination of a collection of random variables following Gaussian distributions still follows a Gaussian distribution. For example, given two random variables $X \sim \mathcal{N}(\mu_X, \sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$, then

$$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2) \tag{14.7}$$

We can generalize this result to higher dimensions.

Definition 14.3. Normal Vector: A random vector $\mathbf{X}$ is said to be normal or Gaussian, if every random variable X_i within it:

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{X_i}, \sigma_{X_i}^2)$$

Definition 14.4. Standard Normal vector: A random vector $\mathbf{Z}$ is said to be normal or Gaussian, if every random variable within it if every random variable Z_i within it:

$$Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$$

Theorem 14.5. A n-dimensional standard Normal vector $\mathbf{Z}$, denoted by, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ has the following join p.d.f.:

$$f_{\mathbf{Z}}(\mathbf{z}) = (\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2}\mathbf{z}^\top\mathbf{z}\right)$$

Proof. Since $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, Their joint p.d.f. is

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) &= (\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} Z_i^2\right) \\ &= (\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2}\mathbf{z}^\top\mathbf{z}\right) \end{aligned}$$

We can verify its expectation and covariance matrix:

$$\boldsymbol{\mu}_{\mathbf{Z}} = \mathbb{E}(\mathbf{Z}) = \mathbf{0}$$

$$\mathbf{\Sigma}_{\mathbf{Z}} = \mathbb{E}(\mathbf{Z}\mathbf{Z}^\top) - \boldsymbol{\mu}_{\mathbf{Z}}\boldsymbol{\mu}_{\mathbf{Z}}^\top = \mathbf{I}$$

$\mathbf{\Sigma}_{\mathbf{Z}} = \mathbf{I}$ is derived from the fact that $\forall i \neq j : \mathbb{E}(Z_iZ_j) = \mathbb{E}(Z_i)\mathbb{E}(Z_j) = 0$ and $Z_i^2 \sim \chi_1^2$ with $\mathbb{E}(\chi_1^2) = 1$. □

Next, we are going to derive the joint p.d.f. of a normal random vector $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{\Sigma}_{\mathbf{X}})$ with $\det(\mathbf{\Sigma}_{\mathbf{X}}) > 0$.

Remark 14.2. Here, we add an assumption, $\det(\mathbf{\Sigma}_{\mathbf{X}}) > 0$, on $\mathbf{X}$. If $\det(\mathbf{\Sigma}_{\mathbf{X}}) = 0$, then it can be shown that some X_i

can be written as a linear combination of the others, so indeed we can remove X_i from the random vector without losing any information.

Since $\mathbf{\Sigma_X}$ is symmetric, by the theorem 9.10

$$\mathbf{\Sigma_X} = \mathbf{PDP}^\top$$

where $\mathbf{P}$ is an orthogonal matrix. $\det(\mathbf{\Sigma_X}) > 0$ guarantees that the diagonal entries of $\mathbf{D}$ is positive, so we can write $\mathbf{D}$ as $\mathbf{D}^{1/2}\mathbf{D}^{1/2}$. Let

$$\mathbf{A} = \mathbf{PD}^{1/2}\mathbf{P}^\top$$

It is easy to check $\mathbf{A}$ is symmetric, non-singular and

$$\mathbf{AA}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{\Sigma_X}$$

Let $\mathbf{Z}$ be a standard Gaussian vector as defined in theorem 14.5 and

$$\mathbf{X} = \mathbf{AZ} + \mathbf{b}$$

Note that $\mathbf{X}$ is also a random vector due to the randomness of $\mathbf{Z}$. We can get

$$\mathbb{E}(\mathbf{X}) = \mathbb{E}(\mathbf{AZ} + \mathbf{b}) = \mathbf{0} + \mathbf{b} = \mathbf{b}$$

$$\mathrm{Var}(\mathbf{X}) = \mathbf{A}\mathrm{Var}(\mathbf{Z})\mathbf{A}^\top = \mathbf{AIA}^\top = \mathbf{\Sigma_X}$$

We can get the joint p.d.f. of $\mathbf{X}$ as in example 14.2 and theorem 14.5:

$$\begin{aligned}
f_\mathbf{X}(\mathbf{x}) &= (\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2}\Big(\mathbf{A}^{-1}(\mathbf{x}-\mathbf{b})\Big)^\top \Big(\mathbf{A}^{-1}(\mathbf{x}-\mathbf{b})\Big)\right)|\det \mathbf{A})|^{-1} \\
&= (\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top(\mathbf{A}^{-1})^\top\mathbf{A}^{-1}(\mathbf{x}-\mathbf{b})\right)|\det \mathbf{A})|^{-1} \\
&= (\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top(\mathbf{AA}^\top)^{-1}(\mathbf{x}-\mathbf{b})\right)|\det \mathbf{A})|^{-1} \\
&= (\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top\mathbf{\Sigma_x}^{-1}(\mathbf{x}-\mathbf{b})\right)|\det \mathbf{A})|^{-1}
\end{aligned}$$

Note that $\det(\mathbf{P})\det(\mathbf{P}^\top) = 1$, since $\mathbf{P}$ is an orthogonal matrix.

$$\det(\mathbf{A}) = \det(\mathbf{P})\det(\mathbf{D}^{1/2})\det(\mathbf{P}^\top) = \sqrt{\det(\mathbf{A})\det(\mathbf{A}^\top)} = \sqrt{\det(\mathbf{\Sigma_X})}$$

By substituting $\det(\mathbf{A}) = \det(\mathbf{\Sigma_X})$ and $\mathbf{b} = \boldsymbol{\mu}_\mathbf{X}$, we can get the following theorem.

Theorem 14.6. A normal vector or Gaussian vector, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \mathbf{\Sigma_X})$ has the following joint p.d.f.:

$$f_\mathbf{X}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\det(\mathbf{\Sigma_X})^{1/2}} \exp\left(\frac{-(\mathbf{x}-\boldsymbol{\mu}_\mathbf{X})^\top\mathbf{\Sigma}_\mathbf{X}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{X})}{2}\right) \quad \forall \mathbf{x} \in \mathbb{R}^n \tag{14.8}$$

where **$\mathbf{\Sigma_X}$ is positive definite.**

Remark 14.3. We have performed a linear transformation,

$$\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}_\mathbf{X} \tag{14.9}$$

on $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and then get a new normal vector $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \mathbf{\Sigma_X})$. Note that

$$\boldsymbol{\mu}_\mathbf{X} = \mathbb{E}(\mathbf{X}) = \mathbf{A}\boldsymbol{\mu}_\mathbf{Z} + \boldsymbol{\mu}_\mathbf{X}$$

$$\mathrm{Var}(\mathbf{X}) = \mathrm{Var}(\mathbf{AZ} + \boldsymbol{\mu}_\mathbf{X}) = \mathbf{A\Sigma_Z A}^\top$$

has illustrated the property of a linear transformation for a normal vector.

Theorem 14.7. Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$ be a p-dimensional normal random vector, $\mathbf{A}$ be an $n \times p$ (where $n \leq p$) real matrix with full row rank, and $\mathbf{b}$ be an n-dimensional real vector, then

$$\mathbf{AX} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_\mathbf{X} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^\top) \tag{14.10}$$

Note that if $n > p$, then $\text{rank}(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}) \leq p$, while $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}$ is an $n \times n$ matrix, implying that $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}$ is singular (not invertible), and so is $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^\top$. We can check if $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^\top$ is a symmetric and positive definite matrix, which is a necessary condition for it to be a valid covariance matrix.

$$(\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^\top)^\top = (\mathbf{APDP}^\top\mathbf{A}^\top)^\top = \mathbf{APDP}^\top\mathbf{A}^\top = \mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^\top \tag{14.11}$$

says the matrix is symmetric. We can verify whether it is positive or not by the definition 10.2. $\forall \mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$,

$$\mathbf{v}^\top\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^\top\mathbf{v} = (\mathbf{A}^\top\mathbf{v})^\top\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^\top\mathbf{v} \tag{14.12}$$

Since $\boldsymbol{\Sigma}_\mathbf{X}$ is positive definite, so is $\mathbf{A}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{A}^\top$. Thus, **$n \leq p$ and $\mathbf{A}$ being of full row rank are two important requirements.**

Theorem 14.8. Suppose a p-dimensional random vector $X \sim \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$. Then $X_i \sim \mathcal{N}(\mu_i, \sigma_{ii})$, where μ_i is the i^{th} element in $\boldsymbol{\mu}_\mathbf{X}$ and σ_{ii} is the i^{th} element of the main diagonal of $\boldsymbol{\Sigma}_\mathbf{X}$.

Proof. Let $\mathbf{e}_i$ be a standard basis vector. Then,

$$X_i = \mathbf{e}_i^\top\mathbf{X} \sim \mathcal{N}(\mathbf{e}_i^\top\boldsymbol{\mu}_\mathbf{X}, \mathbf{e}_i^\top\boldsymbol{\Sigma}_\mathbf{X}\mathbf{e}_i) \tag{14.13}$$

$\square$

Generally, we cannot say that the two random variables X, Y are independent if $\text{Cov}(X, Y) = 0$, except X, Y are normally distributed.

Theorem 14.9. Suppose X, Y are two normal random variables with $\text{Cov}(X, Y) = 0$, then X, Y are independent.

Proof. Let $\mathbf{S} = \begin{bmatrix}\mathbf{X} & \mathbf{Y}\end{bmatrix}^\top$, then $\boldsymbol{\Sigma}_\mathbf{S}$ is a diagonal matrix. By using this fact, expanding the theorem 14.6 can get $f_\mathbf{S}(\mathbf{s}) = f_X(x)f_Y(y)$. $\square$

14.4 Equivalent Representations in a Normal Linear Regression Model

A p-dimensional normal vector $\mathbf{X} \sim \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$ is a convenient way to represent a set of mutually independent random variables, where $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. By the theorem 14.6, we can get

$$f_\mathbf{X}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\prod_{i=1}^{p}\sigma_i}\exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{(X_i - \mu_i)^2}{\sigma_i^2}\right)$$

where

$$(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^\top\boldsymbol{\Sigma}_\mathbf{X}^{-1}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) = \sum_{i=1}^{p}\frac{(X_i - \mu_i)^2}{\sigma_i^2} \tag{14.14}$$

and $\det(\boldsymbol{\Sigma}_\mathbf{X})^{1/2} = \prod_{i=}^{p}\sigma_i$. Thus,

$$X_i \overset{\text{independent}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2) \Longleftrightarrow \mathbf{X} \sim \mathcal{N}\Big(\boldsymbol{\mu}_\mathbf{X}, \text{diag}(\sigma_1^2, \cdots, \sigma_p^2)\Big) \tag{14.15}$$

Definition 14.5. A **Normal Linear Regression Model** is defined as below

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where } \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{14.16}$$

where $\mathbf{X}$ is a full-column-rank $n \times p$ ($n \geq p$) real matrix with $\mathbb{1}$ (a vector with all 1's) as its first column, and $\boldsymbol{\beta} \in \mathbb{R}^p$. The following statements are equivalent:

1. $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$
2. $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$
3. $Y_i \overset{\text{independent}}{\sim} \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$
4. $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$

14.5 Standardizing a Normal Vector

Suppose $\boldsymbol{\Sigma}_\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{P}^\top$ is positive definite. If we let $\mathbf{A} = \mathbf{P}\mathbf{D}^{1/2}\mathbf{P}^\top$, it is clear that

$$\boldsymbol{\Sigma}_\mathbf{X} = \mathbf{A}\mathbf{A} = \mathbf{A}^2 \tag{14.17}$$

Therefore, we can define

$$\boldsymbol{\Sigma}_\mathbf{X}^{1/2} = \mathbf{P}\mathbf{D}^{1/2}\mathbf{P}^\top = (\boldsymbol{\Sigma}_\mathbf{X}^{1/2})^\top \tag{14.18}$$

Note that **$\boldsymbol{\Sigma}_\mathbf{X}^{1/2}$ is symmetric, non-singular, and still positive definite** due to the positive definiteness of $\boldsymbol{\Sigma}_\mathbf{X}$, that is, there is no zero entry on the main diagonal of $\mathbf{D}$. It is easy to check that $\boldsymbol{\Sigma}_\mathbf{X}^{1/2}$ has the following property:

$$\boldsymbol{\Sigma}_\mathbf{X}^{1/2} = \boldsymbol{\Sigma}_\mathbf{X}^{1/2}\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Sigma}_\mathbf{X}\boldsymbol{\Sigma}_\mathbf{X}^{1/2} \tag{14.19}$$

If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$, we can get a standard normal vector by letting

$$\mathbf{Z} = \boldsymbol{\Sigma}_\mathbf{X}{}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X}) \tag{14.20}$$

It is easy to verify that $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{Z}) = \mathbf{I}$.

14.6 The Distribution of LSE

In a linear model, we have found an estimator according to the definition 6.1:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

given that $\mathbf{X}^\top\mathbf{X}$ is non-singular.

Theorem 14.10. Suppose, in a linear model, the response vector $\mathbf{Y}$ has $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Var}(\mathbf{Y}) = \boldsymbol{\Sigma}$. Then the LSE $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ has the following properties:

1. $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
2. $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$

Note that the property 2 uses the fact that the inverse of a symmetric matrix is also symmetric. Note also that **$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}$ if the model is a normal linear model as discussed in definition 14.5.**

Example 14.3. Suppose $\text{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}$. Show that $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

Proof.

$$\begin{aligned}\text{Var}(\widehat{\boldsymbol{\beta}}) &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\end{aligned}$$

□

Example 14.4. Suppose $Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Find the distribution of $\widehat{\boldsymbol{\beta}}$.

Solution. Since $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$. By theorem 14.7,

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

14.7 Estimation of σ^2

Under the assumption of a linear model, we see that $\mathbb{E}(Y_i) = \beta_0 + \beta_i x_{i1} + \cdots + \beta_i x_{ik} = \mathbf{x}_i^\top\boldsymbol{\beta}$, where $\mathbf{x}_i^\top$ is the i^{th} column of $\mathbf{X}$, and $\text{Var}(Y_i) = \sigma^2 = \mathbb{E}\Big((Y_i - \mathbb{E}(Y_i))^2\Big) = \mathbb{E}\Big((Y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2\Big)$. However, $\boldsymbol{\beta}$ is unknown. Intuitively, we can use $\widehat{\boldsymbol{\beta}}$ to estimate σ^2.

Definition 14.6. We can estimate σ^2 by a corresponding average from the sample

$$s^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})^2 = \frac{\text{SSE}(\mathbf{Y})}{n-p-1} \tag{14.21}$$

where n is the sample size and p is the number of x_i's.

Remark 14.4. Here, the design matrix $\mathbf{X}$ is an $n \times (p+1)$ matrix with $\mathbb{1}$ as its first column.

Theorem 14.11. s^2 defined in definition 14.6 is an unbiased estimator of σ^2.

Proof. Given $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}$, by applying the property 8 in theorem 8.1,

$$\begin{aligned}E(\mathbf{Y}^\top(\mathbf{I}-\mathbf{H})\mathbf{Y}) &= \text{tr}\Big((\mathbf{I}-\mathbf{H})\sigma^2\Big) + (\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2\Big(n - \text{tr}(\mathbf{H})\Big) + \mathbf{0} \quad \text{Since } \mathbf{I}-\mathbf{H} \text{ is the orthogonal projection matrix of } \text{Col}(\mathbf{X}) \\ &= \sigma^2\big(n - \text{tr}\Big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\big) = \sigma^2 n - \text{tr}\Big(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big) \\ &= \sigma^2\Big(n - \text{tr}(\mathbf{I}_{p+1})\Big) \\ &= \sigma^2(n-p-1)\end{aligned}$$

Hence, $\mathbb{E}(SSE) = \sigma^2(n-p-1)$. □

Theorem 14.12. In a **normal** linear model, we can find an unbiased estimator for $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = s^2(\mathbf{X}^\top\mathbf{X})^{-1} \tag{14.22}$$

Proof. We know that $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}$,

$$\mathbb{E}(s^2\mathbf{I}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \mathrm{Var}(\widehat{\boldsymbol{\beta}}) \tag{14.23}$$

□