

Linear Algebra & Statistics: Foundations Revisited

A Refined Perspective

April 2, 2025

Abstract

These notes revisit fundamental concepts in linear algebra and their application in statistics, focusing on vector spaces, bases, linear transformations, eigenvalues, eigenvectors, spectral decomposition, projection matrices, and vector calculus. The aim is to provide a clear, intuitive, and rigorous understanding of these essential tools.

1 Vector Spaces, Bases, and Coordinates

Linear algebra provides the language to describe and manipulate objects in spaces equipped with addition and scalar multiplication. Central to this is the idea of a basis, which allows us to coordinatize these spaces.

Lemma 1.1 (Matrix-Vector Product as Linear Combination). Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{u} \in \mathbb{R}^m$. The product $\mathbf{v} = \mathbf{A}\mathbf{u}$ is a linear combination of the columns of \mathbf{A} . Specifically, if $\mathbf{A}^1, \dots, \mathbf{A}^m$ are the columns of \mathbf{A} and $\mathbf{u} = (u_1, \dots, u_m)^T$, then

$$\mathbf{v} = \sum_{j=1}^m u_j \mathbf{A}^j$$

Consequently, \mathbf{v} belongs to the column space of \mathbf{A} , denoted $\text{colspace}(\mathbf{A})$.

Proof. The i -th component of \mathbf{v} is given by $v_i = \sum_{j=1}^m A_{ij}u_j$. Let $[\mathbf{A}^j]_i$ denote the i -th component of the j -th column vector \mathbf{A}^j . Since $A_{ij} = [\mathbf{A}^j]_i$, we have $v_i = \sum_{j=1}^m [\mathbf{A}^j]_i u_j$. This is precisely the i -th component of the sum $\sum_{j=1}^m u_j \mathbf{A}^j$. Therefore, $\mathbf{v} = \sum_{j=1}^m u_j \mathbf{A}^j$. \square

Definition 1.1 (Coordinate Vector). Let \mathcal{V} be a vector space and $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ be an ordered basis for \mathcal{V} . Any vector $\mathbf{v} \in \mathcal{V}$ can be uniquely expressed as a linear combination of the basis vectors:

$$\mathbf{v} = \sum_{i=1}^n k_i \mathbf{b}_i$$

The vector $[\mathbf{v}]_{\mathcal{B}} = (k_1, \dots, k_n)^T \in \mathbb{R}^n$ is called the **coordinate vector** of \mathbf{v} with respect to the basis \mathcal{B} . If we arrange the basis vectors as columns of a matrix $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_n]$, then the relationship can be written compactly as $\mathbf{v} = \mathbf{B}[\mathbf{v}]_{\mathcal{B}}$.

Remark 1.1 (Standard Basis). If $\mathcal{V} = \mathbb{R}^n$ and $\mathcal{B} = \mathcal{E} = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is the standard basis (where \mathbf{e}_i has a 1 in the i -th position and 0s elsewhere), then the coordinate vector $[\mathbf{v}]_{\mathcal{E}}$ is simply the vector \mathbf{v} itself. Thinking in terms of coordinate systems, the standard basis corresponds to the familiar Cartesian axes. Choosing a different basis is akin to choosing a different set of reference axes.

Example 1.1 (Finding Coordinates). Let $\mathcal{V} = \mathbb{R}^2$, $\mathbf{v} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$, and consider the basis $\mathcal{B} = \left\{ \mathbf{b}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{b}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$. We seek scalars k_1, k_2 such that $\mathbf{v} = k_1 \mathbf{b}_1 + k_2 \mathbf{b}_2$. This translates to the matrix equation:

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$$

Solving this system (e.g., using Gaussian elimination on the augmented matrix):

$$\left[\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & 5 \end{array} \right] \xrightarrow{R_1 \leftarrow R_1 - R_2} \left[\begin{array}{cc|c} 1 & 0 & -3 \\ 0 & 1 & 5 \end{array} \right]$$

Thus, $k_1 = -3$ and $k_2 = 5$. The coordinate vector is $[\mathbf{v}]_{\mathcal{B}} = \begin{pmatrix} -3 \\ 5 \end{pmatrix}$. We can verify: $-3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 5 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \end{pmatrix} + \begin{pmatrix} 5 \\ 5 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix} = \mathbf{v}$.

2 Linear Transformations and Change of Basis

Linear transformations preserve the structure of vector spaces. Representing them with matrices depends crucially on the chosen bases for the domain and codomain.

Definition 2.1 (Change-of-Basis Matrix). Let $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ and $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ be two ordered bases for a vector space \mathcal{V} . The **change-of-basis matrix** from \mathcal{B} to \mathcal{C} , denoted $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}$, is the unique matrix such that for any $\mathbf{v} \in \mathcal{V}$:

$$[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}} [\mathbf{v}]_{\mathcal{B}} = [\mathbf{v}]_{\mathcal{C}}$$

This matrix transforms coordinates relative to \mathcal{B} into coordinates relative to \mathcal{C} .

Proposition 2.1 (Structure of Change-of-Basis Matrix). Let $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_n]$ and $\mathbf{C} = [\mathbf{c}_1 \dots \mathbf{c}_n]$ be matrices whose columns are the basis vectors of \mathcal{B} and \mathcal{C} respectively (expressed in some common basis, typically the standard basis if $\mathcal{V} = \mathbb{R}^n$). Then:

$$[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}} = \mathbf{C}^{-1} \mathbf{B}$$

The j -th column of $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}$ is the coordinate vector of the j -th basis vector of \mathcal{B} relative to the basis \mathcal{C} , i.e., $[\mathbf{I}]_{\mathcal{C},j}^{\mathcal{B}} = [\mathbf{b}_j]_{\mathcal{C}}$.

Proof. We know $\mathbf{v} = \mathbf{B}[\mathbf{v}]_{\mathcal{B}}$ and $\mathbf{v} = \mathbf{C}[\mathbf{v}]_{\mathcal{C}}$. Thus, $\mathbf{C}[\mathbf{v}]_{\mathcal{C}} = \mathbf{B}[\mathbf{v}]_{\mathcal{B}}$. Since \mathbf{C} is invertible (its columns form a basis), we have $[\mathbf{v}]_{\mathcal{C}} = \mathbf{C}^{-1} \mathbf{B}[\mathbf{v}]_{\mathcal{B}}$. Comparing this with the definition $[\mathbf{v}]_{\mathcal{C}} = [\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}} [\mathbf{v}]_{\mathcal{B}}$, we get $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}} = \mathbf{C}^{-1} \mathbf{B}$. For the second part, let $\mathbf{v} = \mathbf{b}_j$. Then $[\mathbf{v}]_{\mathcal{B}} = \mathbf{e}_j$ (the j -th standard basis vector). The equation becomes $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}} \mathbf{e}_j = [\mathbf{b}_j]_{\mathcal{C}}$. But $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}} \mathbf{e}_j$ is precisely the j -th column of $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}$. \square

Proposition 2.2 (Inverse Change-of-Basis). The change-of-basis matrix from \mathcal{C} to \mathcal{B} is the inverse of the matrix from \mathcal{B} to \mathcal{C} :

$$[\mathbf{I}]_{\mathcal{B}}^{\mathcal{C}} = ([\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}})^{-1}$$

Proof. Follows directly from $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}} = \mathbf{C}^{-1} \mathbf{B}$ and $[\mathbf{I}]_{\mathcal{B}}^{\mathcal{C}} = \mathbf{B}^{-1} \mathbf{C}$. \square

Remark 2.1. If \mathcal{B} is a basis for \mathbb{R}^n represented by the matrix \mathbf{B} , then $\mathbf{B} = [\mathbf{I}]_{\mathcal{E}}^{\mathcal{B}}$ is the change-of-basis matrix from \mathcal{B} to the standard basis \mathcal{E} . Its inverse, $\mathbf{B}^{-1} = [\mathbf{I}]_{\mathcal{B}}^{\mathcal{E}}$, is the change-of-basis matrix from the standard basis \mathcal{E} to \mathcal{B} .

Definition 2.2 (Matrix Representation of a Linear Transformation). Let $T : \mathcal{V} \rightarrow \mathcal{W}$ be a linear transformation, $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ a basis for \mathcal{V} , and $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_m\}$ a basis for \mathcal{W} . The **matrix representation** of T with respect to bases \mathcal{B} and \mathcal{C} , denoted $[T]_{\mathcal{C}}^{\mathcal{B}}$, is the unique $m \times n$ matrix such that for all $\mathbf{v} \in \mathcal{V}$:

$$[T]_{\mathcal{C}}^{\mathcal{B}}[\mathbf{v}]_{\mathcal{B}} = [T(\mathbf{v})]_{\mathcal{C}}$$

The j -th column of $[T]_{\mathcal{C}}^{\mathcal{B}}$ is the coordinate vector of $T(\mathbf{b}_j)$ relative to the basis \mathcal{C} , i.e., $[T]_{\mathcal{C},j}^{\mathcal{B}} = [T(\mathbf{b}_j)]_{\mathcal{C}}$.

Remark 2.2. The change-of-basis matrix $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}$ is a special case where T is the identity transformation $I : \mathcal{V} \rightarrow \mathcal{V}$.

Theorem 2.1 (Changing Bases for a Linear Transformation). Let $T : \mathcal{V} \rightarrow \mathcal{V}$ be a linear transformation, and let \mathcal{B}, \mathcal{C} be two bases for \mathcal{V} . Let $[T]_{\mathcal{B}}$ be the matrix representation of T relative to \mathcal{B} (i.e., $[T]_{\mathcal{B}}^{\mathcal{B}}$) and $[T]_{\mathcal{C}}$ be the matrix representation relative to \mathcal{C} . Then these matrices are related by similarity transformation:

$$[T]_{\mathcal{C}} = [\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}[T]_{\mathcal{B}}[\mathbf{I}]_{\mathcal{B}}^{\mathcal{C}}$$

where $[\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}$ is the change-of-basis matrix from \mathcal{B} to \mathcal{C} , and $[\mathbf{I}]_{\mathcal{B}}^{\mathcal{C}} = ([\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}})^{-1}$ is the change-of-basis matrix from \mathcal{C} to \mathcal{B} .

Proof. For any $\mathbf{v} \in \mathcal{V}$, we want to relate $[T(\mathbf{v})]_{\mathcal{C}}$ to $[\mathbf{v}]_{\mathcal{C}}$. We know:

$$\begin{aligned} [T(\mathbf{v})]_{\mathcal{C}} &= [T]_{\mathcal{C}}[\mathbf{v}]_{\mathcal{C}} \\ [T(\mathbf{v})]_{\mathcal{C}} &= [\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}[T(\mathbf{v})]_{\mathcal{B}} \quad (\text{Change basis of the output}) \\ &= [\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}([T]_{\mathcal{B}}[\mathbf{v}]_{\mathcal{B}}) \quad (\text{Apply } T \text{ in basis } \mathcal{B}) \\ &= [\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}[T]_{\mathcal{B}}([\mathbf{I}]_{\mathcal{B}}^{\mathcal{C}}[\mathbf{v}]_{\mathcal{C}}) \quad (\text{Change basis of the input}) \end{aligned}$$

So, $[T]_{\mathcal{C}}[\mathbf{v}]_{\mathcal{C}} = ([\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}[T]_{\mathcal{B}}[\mathbf{I}]_{\mathcal{B}}^{\mathcal{C}})[\mathbf{v}]_{\mathcal{C}}$. Since this holds for all \mathbf{v} , the matrices must be equal: $[T]_{\mathcal{C}} = [\mathbf{I}]_{\mathcal{C}}^{\mathcal{B}}[T]_{\mathcal{B}}[\mathbf{I}]_{\mathcal{B}}^{\mathcal{C}}$. \square

Example 2.1 (Applying the Change of Basis Formula). Consider the linear transformation $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $F(x, y) = (5x - y, 2x + y)$. The standard basis is $\mathcal{E} = \{\mathbf{e}_1 = (1, 0), \mathbf{e}_2 = (0, 1)\}$. The matrix representation of F relative to \mathcal{E} is found by applying F to the basis vectors: $F(\mathbf{e}_1) = F(1, 0) = (5, 2) = 5\mathbf{e}_1 + 2\mathbf{e}_2$, $F(\mathbf{e}_2) = F(0, 1) = (-1, 1) = -1\mathbf{e}_1 + 1\mathbf{e}_2$. So, $\mathbf{A} = [F]_{\mathcal{E}} = \begin{pmatrix} 5 & -1 \\ 2 & 1 \end{pmatrix}$.

Now consider a different basis $S = \{\mathbf{u}_1 = (1, 4), \mathbf{u}_2 = (2, 7)\}$. We want to find the matrix $\mathbf{B} = [F]_S$. The change-of-basis matrix from S to \mathcal{E} is simply the matrix whose columns are the vectors in S : $\mathbf{P} = [\mathbf{I}]_{\mathcal{E}}^S = \begin{pmatrix} 1 & 2 \\ 4 & 7 \end{pmatrix}$. The change-of-basis matrix from \mathcal{E} to S is its inverse:

$$\mathbf{P}^{-1} = [\mathbf{I}]_S^{\mathcal{E}} = \frac{1}{(1)(7) - (2)(4)} \begin{pmatrix} 7 & -2 \\ -4 & 1 \end{pmatrix} = \begin{pmatrix} -7 & 2 \\ 4 & -1 \end{pmatrix}.$$

Using the theorem:

$$\begin{aligned}
\mathbf{B} &= [F]_S = [\mathbf{I}]_S^\mathcal{E} [F]_\mathcal{E} [\mathbf{I}]_\mathcal{E}^S \\
&= \mathbf{P}^{-1} \mathbf{A} \mathbf{P} \\
&= \begin{pmatrix} -7 & 2 \\ 4 & -1 \end{pmatrix} \begin{pmatrix} 5 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 4 & 7 \end{pmatrix} \\
&= \begin{pmatrix} -35 + 4 & 7 + 2 \\ 20 - 2 & -4 - 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 4 & 7 \end{pmatrix} \\
&= \begin{pmatrix} -31 & 9 \\ 18 & -5 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 4 & 7 \end{pmatrix} \\
&= \begin{pmatrix} -31 + 36 & -62 + 63 \\ 18 - 20 & 36 - 35 \end{pmatrix} \\
&= \begin{pmatrix} 5 & 1 \\ -2 & 1 \end{pmatrix}
\end{aligned}$$

So, the matrix representing F in the basis S is $\mathbf{B} = \begin{pmatrix} 5 & 1 \\ -2 & 1 \end{pmatrix}$.

Remark 2.3 (Default Basis). Unless specified otherwise, a matrix \mathbf{A} representing a linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is assumed to be with respect to the standard bases in both the domain and codomain, i.e., $\mathbf{A} = [T]_\mathcal{E}^\mathcal{E}$.

3 Kernel, Image, and the Rank-Nullity Theorem

The kernel and image are fundamental subspaces associated with a linear transformation (and its matrix representation).

Definition 3.1 (Kernel and Image). Let $T : \mathcal{V} \rightarrow \mathcal{W}$ be a linear transformation, represented by matrix \mathbf{A} (typically w.r.t. standard bases).

- The **Kernel** (or Null Space) of T is the set of vectors in \mathcal{V} that map to the zero vector in \mathcal{W} :

$$\text{Ker}(T) = \{\mathbf{v} \in \mathcal{V} \mid T(\mathbf{v}) = \mathbf{0}\}$$

In terms of the matrix \mathbf{A} , this corresponds to the null space of \mathbf{A} :

$$\text{Ker}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0}\}$$

Finding a basis for $\text{Ker}(T)$ involves finding a basis for the solution space of the homogeneous system $\mathbf{A}\mathbf{x} = \mathbf{0}$.

- The **Image** (or Range) of T is the set of all possible outputs in \mathcal{W} :

$$\text{Im}(T) = \{T(\mathbf{v}) \mid \mathbf{v} \in \mathcal{V}\} = \{\mathbf{w} \in \mathcal{W} \mid \exists \mathbf{v} \in \mathcal{V}, T(\mathbf{v}) = \mathbf{w}\}$$

In terms of the matrix \mathbf{A} , this corresponds to the column space of \mathbf{A} :

$$\text{Im}(\mathbf{A}) = \text{colspace}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\} = \{\mathbf{w} \in \mathbb{R}^m \mid \exists \mathbf{x} \in \mathbb{R}^n, \mathbf{A}\mathbf{x} = \mathbf{w}\}$$

A basis for $\text{Im}(T)$ can be found by finding a basis for the column space of \mathbf{A} . A common method is to find the row space of \mathbf{A}^T (since $\text{colspace}(\mathbf{A}) = \text{rowspace}(\mathbf{A}^T)$) by row-reducing \mathbf{A}^T and taking the non-zero rows.

Theorem 3.1 (Rank-Nullity Theorem). For any linear transformation $T : \mathcal{V} \rightarrow \mathcal{W}$, where \mathcal{V} is finite-dimensional:

$$\dim(\text{Ker}(T)) + \dim(\text{Im}(T)) = \dim(\mathcal{V})$$

In terms of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ representing T :

$$\dim(\text{Ker}(\mathbf{A})) + \dim(\text{Im}(\mathbf{A})) = n$$

where $\dim(\text{Ker}(\mathbf{A}))$ is the nullity of \mathbf{A} , $\dim(\text{Im}(\mathbf{A}))$ is the rank of \mathbf{A} (which equals the rank of T), and n is the dimension of the domain (number of columns in \mathbf{A}). If $\text{rank}(\mathbf{A}) = r$, then $\dim(\text{Im}(\mathbf{A})) = r$ and $\dim(\text{Ker}(\mathbf{A})) = n - r$. Note that $r \leq \min(n, m)$.

4 Eigenvalues, Eigenvectors, and Spectral Decomposition

Eigenvalues and eigenvectors reveal the intrinsic geometry of a linear transformation – directions that are simply scaled by the transformation.

Definition 4.1 (Eigenvalue and Eigenvector). Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a square matrix. A scalar $\lambda \in \mathbb{C}$ (or \mathbb{R}) is an **eigenvalue** of \mathbf{A} if there exists a non-zero vector $\mathbf{x} \in \mathbb{C}^n$ (or \mathbb{R}^n) such that:

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

The vector \mathbf{x} is called an **eigenvector** corresponding to the eigenvalue λ .

Definition 4.2 (Characteristic Polynomial). The eigenvalues λ are the solutions to the equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, or $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$. For a non-zero eigenvector \mathbf{x} to exist, the matrix $(\mathbf{A} - \lambda\mathbf{I})$ must be singular (non-invertible). This means its determinant must be zero:

$$p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

The polynomial $p(\lambda)$ is called the **characteristic polynomial** of \mathbf{A} . Its roots are the eigenvalues of \mathbf{A} .

Theorem 4.1 (Spectral Theorem for Symmetric Matrices). Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix ($\mathbf{A} = \mathbf{A}^T$). Then:

1. All eigenvalues of \mathbf{A} are real.
2. Eigenvectors corresponding to distinct eigenvalues are orthogonal.
3. \mathbf{A} is orthogonally diagonalizable. That is, there exists an orthogonal matrix \mathbf{U} (whose columns form an orthonormal basis of eigenvectors for \mathbb{R}^n , so $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$) and a diagonal matrix $\mathbf{\Lambda}$ (whose diagonal entries are the eigenvalues of \mathbf{A}) such that:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

This is called the **spectral decomposition** of \mathbf{A} . Equivalently, $\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$.

Remark 4.1. The columns of \mathbf{U} are the orthonormal eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_n$, and the diagonal entries of $\mathbf{\Lambda}$ are the corresponding eigenvalues $\lambda_1, \dots, \lambda_n$. The spectral decomposition expresses \mathbf{A} as a sum of rank-one projection matrices scaled by the eigenvalues: $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^T$.

Example 4.1 (Interpretation of Spectral Decomposition). Consider $\mathbf{A} = \mathbf{X}^T\mathbf{X} \in \mathbb{R}^{p \times p}$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$. Since \mathbf{A} is symmetric, it has a spectral decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. Assume the eigenvalues $\lambda_1, \dots, \lambda_p$ are distinct. The matrix \mathbf{A} represents the linear transformation $T(\mathbf{v}) = \mathbf{A}\mathbf{v}$ with respect to the standard basis \mathcal{E} . However, consider the basis $\mathcal{B} = \{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ formed by the orthonormal eigenvectors (columns of \mathbf{U}). What is the matrix representation of T with respect to \mathcal{B} ? Let $[T]_{\mathcal{B}}$ be this matrix. By the change of basis formula:

$$[T]_{\mathcal{B}} = [\mathbf{I}]_{\mathcal{B}}^{\mathcal{E}} [T]_{\mathcal{E}} [\mathbf{I}]_{\mathcal{E}}^{\mathcal{B}}$$

Here, $[T]_{\mathcal{E}} = \mathbf{A}$. The change-of-basis matrix from \mathcal{B} to \mathcal{E} is $\mathbf{U} = [\mathbf{I}]_{\mathcal{E}}^{\mathcal{B}}$. The change-of-basis matrix from \mathcal{E} to \mathcal{B} is $\mathbf{U}^{-1} = \mathbf{U}^T = [\mathbf{I}]_{\mathcal{B}}^{\mathcal{E}}$. So,

$$[T]_{\mathcal{B}} = \mathbf{U}^T \mathbf{A} \mathbf{U} = \mathbf{U}^T (\mathbf{U} \mathbf{A} \mathbf{U}^T) \mathbf{U} = (\mathbf{U}^T \mathbf{U}) \mathbf{A} (\mathbf{U}^T \mathbf{U}) = \mathbf{I} \mathbf{A} \mathbf{I} = \mathbf{A}$$

This confirms that the matrix representing the transformation $T(\mathbf{v}) = \mathbf{A}\mathbf{v}$ in the basis of eigenvectors \mathcal{B} is the diagonal matrix $\mathbf{\Lambda}$ of eigenvalues. The transformation simply scales the components along the eigenvector directions.

Now consider the transformation $f : \mathbb{R}^p \rightarrow \mathbb{R}^p$ defined by $f(\mathbf{v}) = (\lambda_1 v_1, \dots, \lambda_p v_p)^T$. This transformation scales the standard basis vectors. Its matrix representation in the standard basis \mathcal{E} is precisely $\mathbf{\Lambda}$. The connection is: The transformation $T(\mathbf{v}) = \mathbf{A}\mathbf{v}$ (represented by \mathbf{A} in \mathcal{E}) behaves like the simpler scaling transformation f (represented by $\mathbf{\Lambda}$ in \mathcal{E}) when viewed in the eigenbasis \mathcal{B} . The matrix \mathbf{A} represents a scaling along the axes defined by its eigenvectors.

Remark 4.2 (Geometric Interpretation via SVD). While the spectral decomposition applies to symmetric matrices, the Singular Value Decomposition (SVD) applies to any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$. It states $\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$, where \mathbf{U} and \mathbf{V} are orthogonal matrices and $\mathbf{\Sigma}$ is a diagonal matrix of singular values. Geometrically, multiplying by \mathbf{A} can be seen as: 1. A rotation/reflection in the domain space (by \mathbf{V}^T). 2. A scaling along the principal axes (by $\mathbf{\Sigma}$). 3. A rotation/reflection in the codomain space (by \mathbf{U}). The images provided likely illustrate this decomposition: \mathbf{V}^T aligns the input vector with axes, $\mathbf{\Sigma}$ scales it, and \mathbf{U} rotates the result. For a symmetric matrix $\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$, the rotations are related ($\mathbf{V} = \mathbf{U}$) and the scaling factors are the eigenvalues (possibly negative, absorbed into rotation).

Proposition 4.1 (Determinant and Trace). For any square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ (even non-diagonalizable) with eigenvalues $\lambda_1, \dots, \lambda_n$ (counted with algebraic multiplicity):

1. $\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i$
2. $\text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i$

If \mathbf{A} is diagonalizable, $\mathbf{A} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^{-1}$, the proof is straightforward using properties of determinant and trace: $\det(\mathbf{A}) = \det(\mathbf{P} \mathbf{\Lambda} \mathbf{P}^{-1}) = \det(\mathbf{P}) \det(\mathbf{\Lambda}) \det(\mathbf{P}^{-1}) = \det(\mathbf{\Lambda}) = \prod \lambda_i$. $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{P} \mathbf{\Lambda} \mathbf{P}^{-1}) = \text{tr}(\mathbf{\Lambda} \mathbf{P}^{-1} \mathbf{P}) = \text{tr}(\mathbf{\Lambda}) = \sum \lambda_i$. The result holds even for non-diagonalizable matrices, but the proof involves Jordan Normal Form or properties of the characteristic polynomial.

5 Orthogonal Projection Matrices

Projection matrices are fundamental in statistics (e.g., linear regression) and numerical analysis.

Definition 5.1 (Idempotent Matrix). A square matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ is **idempotent** if $\mathbf{P}^2 = \mathbf{P}$.

Definition 5.2 (Orthogonal Projection Matrix). A matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ is an **orthogonal projection matrix** if it is both:

1. Symmetric: $\mathbf{P}^T = \mathbf{P}$
2. Idempotent: $\mathbf{P}^2 = \mathbf{P}$

Such a matrix projects vectors orthogonally onto its column space (image).

Proposition 5.1 (Eigenvalues of Projection Matrices). The eigenvalues of an orthogonal projection matrix \mathbf{P} are either 0 or 1.

Proof. Let λ be an eigenvalue with eigenvector $\mathbf{x} \neq \mathbf{0}$. So $\mathbf{P}\mathbf{x} = \lambda\mathbf{x}$. Applying \mathbf{P} again: $\mathbf{P}^2\mathbf{x} = \mathbf{P}(\lambda\mathbf{x}) = \lambda(\mathbf{P}\mathbf{x}) = \lambda(\lambda\mathbf{x}) = \lambda^2\mathbf{x}$. Since \mathbf{P} is idempotent, $\mathbf{P}^2 = \mathbf{P}$, so $\mathbf{P}^2\mathbf{x} = \mathbf{P}\mathbf{x} = \lambda\mathbf{x}$. Thus, $\lambda^2\mathbf{x} = \lambda\mathbf{x}$, which implies $(\lambda^2 - \lambda)\mathbf{x} = \mathbf{0}$. Since $\mathbf{x} \neq \mathbf{0}$, we must have $\lambda^2 - \lambda = 0$, or $\lambda(\lambda - 1) = 0$. The only possible eigenvalues are $\lambda = 0$ or $\lambda = 1$. The multiplicity of the

eigenvalue 1 is the dimension of the eigenspace $E_1 = \{\mathbf{x} \mid \mathbf{P}\mathbf{x} = \mathbf{x}\}$, which is the subspace onto which \mathbf{P} projects, i.e., $\text{Im}(\mathbf{P})$. So, multiplicity of 1 is $\text{rank}(\mathbf{P})$. The multiplicity of the eigenvalue 0 is the dimension of the eigenspace $E_0 = \{\mathbf{x} \mid \mathbf{P}\mathbf{x} = \mathbf{0}\}$, which is $\text{Ker}(\mathbf{P})$. By Rank-Nullity, this multiplicity is $n - \text{rank}(\mathbf{P})$. \square

Proposition 5.2 (Projection onto Column Space). Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a matrix with linearly independent columns (full column rank, $p \leq n$). The matrix

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

is the orthogonal projection matrix onto the column space of \mathbf{X} , $\text{colspace}(\mathbf{X}) = \text{Im}(\mathbf{X})$.

Proof. We verify the properties:

1. Symmetry:

$$\begin{aligned}\mathbf{P}_{\mathbf{X}}^T &= (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= (\mathbf{X}^T)^T((\mathbf{X}^T\mathbf{X})^{-1})^T\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad (\text{since } (\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1} \text{ and } \mathbf{X}^T\mathbf{X} \text{ is symmetric}) \\ &= \mathbf{P}_{\mathbf{X}}\end{aligned}$$

2. Idempotence:

$$\begin{aligned}\mathbf{P}_{\mathbf{X}}^2 &= (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{P}_{\mathbf{X}}\end{aligned}$$

3. Image: For any $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{P}_{\mathbf{X}}\mathbf{v} = \mathbf{X}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{v}]$. Let $\mathbf{b}\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{v}$. Then $\mathbf{P}_{\mathbf{X}}\mathbf{v} = \mathbf{X}\mathbf{b}\beta$, which is clearly a linear combination of the columns of \mathbf{X} , hence $\mathbf{P}_{\mathbf{X}}\mathbf{v} \in \text{Im}(\mathbf{X})$. \square

Remark 5.1 (Properties of $\mathbf{P}_{\mathbf{X}}$). The provided list (Proposition 4, 6, 7 in the original notes) summarizes key properties:

- $\mathbf{P}_{\mathbf{X}}$ is symmetric and idempotent (proved above).
- $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$: Projecting the columns of \mathbf{X} onto their own span leaves them unchanged. $(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{I} = \mathbf{X})$
- $\mathbf{X}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \mathbf{0}$: The residual matrix $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})$ projects onto the orthogonal complement of $\text{Im}(\mathbf{X})$. The columns of \mathbf{X} are orthogonal to this space. $(\mathbf{X}^T - \mathbf{X}^T\mathbf{P}_{\mathbf{X}} = \mathbf{X}^T - \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}^T - \mathbf{X}^T = \mathbf{0})$
- $\text{Im}(\mathbf{P}_{\mathbf{X}}) = \text{Im}(\mathbf{X})$ (proved above).
- If $m = n$ and \mathbf{X} is invertible, $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}\mathbf{X}^{-1}(\mathbf{X}^T)^{-1}\mathbf{X}^T = \mathbf{I}\mathbf{I} = \mathbf{I}$. (Projection onto the whole space \mathbb{R}^n).
- $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{v} \in \text{Im}(\mathbf{X})^\perp$: The vector $\mathbf{v} - \mathbf{P}_{\mathbf{X}}\mathbf{v}$ is the residual, orthogonal to the projection space.

- If $\mathbf{w} \in \text{Im}(\mathbf{X})$, then $\mathbf{P}_\mathbf{X}\mathbf{w} = \mathbf{w}$. (Vectors already in the space are unchanged by projection).
- If $\mathbf{w} \in \text{Im}(\mathbf{X})^\perp$, then $\mathbf{P}_\mathbf{X}\mathbf{w} = \mathbf{0}$. (Vectors orthogonal to the space project to the origin).
- $\mathbf{P}_\mathbf{X}$ depends only on the subspace $\text{Im}(\mathbf{X})$, not the specific basis \mathbf{X} . If $\text{Im}(\mathbf{Z}) = \text{Im}(\mathbf{X})$, then $\mathbf{P}_\mathbf{Z} = \mathbf{P}_\mathbf{X}$.
- If $L \subseteq M$ are subspaces, then $\mathbf{P}_M\mathbf{P}_L = \mathbf{P}_L\mathbf{P}_M = \mathbf{P}_L$. (Projecting onto L then M is the same as just projecting onto L).
- $\mathbf{I} - \mathbf{P}_\mathbf{X} = \mathbf{P}_{\text{Im}(\mathbf{X})^\perp}$. The matrix $\mathbf{I} - \mathbf{P}_\mathbf{X}$ projects onto the orthogonal complement of $\text{Im}(\mathbf{X})$.
- If $L \subseteq M$, then $\mathbf{P}_M - \mathbf{P}_L = \mathbf{P}_{M \cap L^\perp}$. (Projects onto the part of M orthogonal to L).
- Any symmetric, idempotent matrix \mathbf{Q} is an orthogonal projection matrix onto its own image, $\text{Im}(\mathbf{Q})$.

6 Orthogonal Complements and Fundamental Subspaces

Definition 6.1 (Orthogonal Complement). Let U be a subspace of an inner product space V . The **orthogonal complement** of U , denoted U^\perp , is the set of all vectors in V that are orthogonal to every vector in U :

$$U^\perp = \{\mathbf{v} \in V \mid \langle \mathbf{u}, \mathbf{v} \rangle = 0 \text{ for all } \mathbf{u} \in U\}$$

In \mathbb{R}^n with the standard dot product, $U^\perp = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{u}^T \mathbf{v} = 0 \text{ for all } \mathbf{u} \in U\}$.

Theorem 6.1 (Direct Sum Decomposition). For any subspace U of a finite-dimensional inner product space V , V is the direct sum of U and its orthogonal complement:

$$V = U \oplus U^\perp$$

This means every $\mathbf{v} \in V$ can be uniquely written as $\mathbf{v} = \mathbf{u} + \mathbf{w}$ where $\mathbf{u} \in U$ and $\mathbf{w} \in U^\perp$. Furthermore, $\mathbf{u} = \text{proj}_U(\mathbf{v})$ and $\mathbf{w} = \text{proj}_{U^\perp}(\mathbf{v})$.

Theorem 6.2 (Fundamental Subspaces Theorem Part 1). For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$:

$$\text{Im}(\mathbf{A}^T) = \text{Ker}(\mathbf{A})^\perp$$

The row space of \mathbf{A} is the orthogonal complement of the null space of \mathbf{A} (subspaces of \mathbb{R}^n). And also:

$$\text{Ker}(\mathbf{A}^T) = \text{Im}(\mathbf{A})^\perp$$

The null space of \mathbf{A}^T (left null space of \mathbf{A}) is the orthogonal complement of the column space of \mathbf{A} (subspaces of \mathbb{R}^m).

Proof of $\text{Im}(\mathbf{A}^T) = \text{Ker}(\mathbf{A})^\perp$. Let $\mathbf{x} \in \text{Ker}(\mathbf{A})$. This means $\mathbf{Ax} = \mathbf{0}$. Let $\mathbf{r}_1, \dots, \mathbf{r}_m$ be the rows of \mathbf{A} . Then $\mathbf{Ax} = \mathbf{0}$ means $\mathbf{r}_i^T \mathbf{x} = 0$ for all $i = 1, \dots, m$. This implies \mathbf{x} is orthogonal to every row of \mathbf{A} . Since the row space $\text{Im}(\mathbf{A}^T)$ is the span of the rows, \mathbf{x} must be orthogonal to every vector in $\text{Im}(\mathbf{A}^T)$. Thus, $\text{Ker}(\mathbf{A}) \subseteq \text{Im}(\mathbf{A}^T)^\perp$.

Now let $\mathbf{y} \in \text{Im}(\mathbf{A}^T)^\perp$. This means \mathbf{y} is orthogonal to every vector in the row space, including the rows themselves: $\mathbf{r}_i^T \mathbf{y} = 0$ for all i . But this is exactly the condition $\mathbf{Ay} = \mathbf{0}$, which means $\mathbf{y} \in \text{Ker}(\mathbf{A})$. Thus, $\text{Im}(\mathbf{A}^T)^\perp \subseteq \text{Ker}(\mathbf{A})$. Combining the inclusions and using the property $(U^\perp)^\perp = U$, we get $\text{Ker}(\mathbf{A})^\perp = (\text{Im}(\mathbf{A}^T)^\perp)^\perp = \text{Im}(\mathbf{A}^T)$. \square

Remark 6.1 (Diagonalizability). A matrix \mathbf{A} is diagonalizable if it has a full set of linearly independent eigenvectors. Not all matrices are diagonalizable. A matrix fails to be diagonalizable if and only if for at least one eigenvalue λ_i , the geometric multiplicity (dimension of the eigenspace $E_{\lambda_i} = \text{Ker}(\mathbf{A} - \lambda_i \mathbf{I})$) is strictly less than the algebraic multiplicity (multiplicity of λ_i as a root of the characteristic polynomial). The original notes mention a condition involving powers $(\mathbf{A} - \lambda_i \mathbf{I})^k = 0$, related to generalized eigenvectors and Jordan Normal Form, which arises when geometric multiplicity $<$ algebraic multiplicity. However, a crucial result is that all symmetric matrices are orthogonally diagonalizable (as stated in the Spectral Theorem).

Corollary 6.1 (Projection onto Orthogonal Complement). Let $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ be the projection onto $\text{Im}(\mathbf{X})$. The matrix $\mathbf{Q} = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$ is the orthogonal projection matrix onto the orthogonal complement space $\text{Im}(\mathbf{X})^\perp$.

Proof. 1. Symmetry: $\mathbf{Q}^T = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})^T = \mathbf{I}^T - \mathbf{P}_{\mathbf{X}}^T = \mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{Q}$. 2. Idempotence: $\mathbf{Q}^2 = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \mathbf{I} - \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}} + \mathbf{P}_{\mathbf{X}}^2 = \mathbf{I} - \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}} + \mathbf{P}_{\mathbf{X}} = \mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{Q}$. Thus \mathbf{Q} is an orthogonal projection matrix. It projects onto $\text{Im}(\mathbf{Q})$. What is $\text{Im}(\mathbf{Q})$? A vector \mathbf{w} is in $\text{Im}(\mathbf{Q})$ if $\mathbf{w} = \mathbf{Q}\mathbf{v} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{v}$ for some \mathbf{v} . We know from property 7 of $\mathbf{P}_{\mathbf{X}}$ that $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{v} \in \text{Im}(\mathbf{X})^\perp$. So $\text{Im}(\mathbf{Q}) \subseteq \text{Im}(\mathbf{X})^\perp$. Conversely, let $\mathbf{y} \in \text{Im}(\mathbf{X})^\perp$. We want to show $\mathbf{y} = \mathbf{Q}\mathbf{y}$. $\mathbf{Q}\mathbf{y} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y} = \mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{y}$. Since $\mathbf{y} \in \text{Im}(\mathbf{X})^\perp$, by property 9, $\mathbf{P}_{\mathbf{X}}\mathbf{y} = \mathbf{0}$. Thus $\mathbf{Q}\mathbf{y} = \mathbf{y}$. This means $\mathbf{y} \in \text{Im}(\mathbf{Q})$. So $\text{Im}(\mathbf{X})^\perp \subseteq \text{Im}(\mathbf{Q})$. Therefore, $\text{Im}(\mathbf{Q}) = \text{Im}(\mathbf{X})^\perp$.

Spectral Decomposition of \mathbf{Q} : Since $\mathbf{P}_{\mathbf{X}}$ is symmetric, it has spectral decomposition $\mathbf{P}_{\mathbf{X}} = \mathbf{U}\mathbf{\Lambda}_P \mathbf{U}^T$, where $\mathbf{\Lambda}_P$ has $r = \text{rank}(\mathbf{P}_{\mathbf{X}})$ ones and $n - r$ zeros on the diagonal. Then $\mathbf{Q} = \mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{U}\mathbf{I}\mathbf{U}^T - \mathbf{U}\mathbf{\Lambda}_P \mathbf{U}^T = \mathbf{U}(\mathbf{I} - \mathbf{\Lambda}_P)\mathbf{U}^T$. The matrix $\mathbf{\Lambda}_Q = \mathbf{I} - \mathbf{\Lambda}_P$ is also diagonal. Its diagonal entries are $1 - 1 = 0$ (where $\mathbf{\Lambda}_P$ had 1) and $1 - 0 = 1$ (where $\mathbf{\Lambda}_P$ had 0). So $\mathbf{\Lambda}_Q$ has $n - r$ ones and r zeros. The eigenvectors (columns of \mathbf{U}) remain the same, but the eigenvalues are flipped (0 becomes 1, 1 becomes 0). The eigenvectors corresponding to eigenvalue 1 for \mathbf{Q} are those corresponding to eigenvalue 0 for $\mathbf{P}_{\mathbf{X}}$ (i.e., a basis for $\text{Ker}(\mathbf{P}_{\mathbf{X}}) = \text{Im}(\mathbf{X})^\perp$). The eigenvectors corresponding to eigenvalue 0 for \mathbf{Q} are those corresponding to eigenvalue 1 for $\mathbf{P}_{\mathbf{X}}$ (i.e., a basis for $\text{Im}(\mathbf{P}_{\mathbf{X}}) = \text{Im}(\mathbf{X})$). \square

Remark 6.2 (Application to OLS). In Ordinary Least Squares (OLS), we want to minimize the squared norm of the residual vector $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$. The solution is found by projecting \mathbf{y} onto the column space of \mathbf{X} . The fitted values are $\hat{\mathbf{y}} = \mathbf{P}_{\mathbf{X}}\mathbf{y}$, and the optimal coefficients are $\hat{\mathbf{b}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. The residual vector is $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{y} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}$. This is the projection of \mathbf{y} onto the orthogonal complement space $\text{Im}(\mathbf{X})^\perp$. The squared norm of the residual is $\|\mathbf{e}\|^2 = \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}\|^2$. Since $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})$ is a projection matrix, this represents the squared distance from \mathbf{y} to the subspace $\text{Im}(\mathbf{X})$. If we increase the rank of \mathbf{X} (by adding more linearly independent predictors), we are projecting onto a larger subspace. Projecting onto a larger subspace can never increase the distance from the original vector, so the residual sum of squares $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$ will generally decrease (or stay the same).

7 Examples and Statistical Connections

Example 7.1 (Rank-One Projection). Let $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v} \neq \mathbf{0}$. Consider the matrix $\mathbf{P} = \frac{\mathbf{v}\mathbf{v}^T}{\|\mathbf{v}\|^2}$. Let's show it's an orthogonal projection matrix. Let $c = \|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$. 1. Symmetry: $\mathbf{P}^T = (c^{-1} \mathbf{v}\mathbf{v}^T)^T = c^{-1} (\mathbf{v}^T)^T \mathbf{v}^T = c^{-1} \mathbf{v}\mathbf{v}^T = \mathbf{P}$. 2. Idempotence: $\mathbf{P}^2 = (c^{-1} \mathbf{v}\mathbf{v}^T)(c^{-1} \mathbf{v}\mathbf{v}^T) = c^{-2} \mathbf{v}(\mathbf{v}^T \mathbf{v})\mathbf{v}^T = c^{-2} \mathbf{v}(c)\mathbf{v}^T = c^{-1} \mathbf{v}\mathbf{v}^T = \mathbf{P}$. So, \mathbf{P} is an orthogonal projection matrix. What does it project onto? Its image is $\text{Im}(\mathbf{P}) = \{\mathbf{P}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\} = \{c^{-1} \mathbf{v}(\mathbf{v}^T \mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}$. Since $\mathbf{v}^T \mathbf{x}$ is a scalar, any output is a scalar multiple of \mathbf{v} . Thus, $\text{Im}(\mathbf{P}) = \text{span}\{\mathbf{v}\}$. It projects onto the line spanned by \mathbf{v} . The rank of the matrix is the dimension of its image, which is

$\dim(\text{span}\{\mathbf{v}\}) = 1$ (since $\mathbf{v} \neq \mathbf{0}$).

Example 7.2 (Sample Variance and Projections). Let Y_1, \dots, Y_n be random variables with mean μ and variance σ^2 . The sample mean is $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. The sample variance is $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. Let $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$. Then $\bar{Y} = \frac{1}{n} \mathbf{1}^T \mathbf{Y}$. The vector of means is $\bar{Y} \mathbf{1} = (\frac{1}{n} \mathbf{1}^T \mathbf{Y}) \mathbf{1} = \frac{\mathbf{1} \mathbf{1}^T}{n} \mathbf{Y}$. Notice that $\mathbf{P}_1 = \frac{\mathbf{1} \mathbf{1}^T}{\|\mathbf{1}\|^2} = \frac{\mathbf{1} \mathbf{1}^T}{n}$ is the projection matrix onto the subspace spanned by $\mathbf{1}$. So, the vector of sample means is $\bar{Y} \mathbf{1} = \mathbf{P}_1 \mathbf{Y}$. The vector of deviations from the mean is $\mathbf{Y} - \bar{Y} \mathbf{1} = \mathbf{Y} - \mathbf{P}_1 \mathbf{Y} = (\mathbf{I} - \mathbf{P}_1) \mathbf{Y}$. The sum of squared deviations is $\|\mathbf{Y} - \bar{Y} \mathbf{1}\|^2 = \|(\mathbf{I} - \mathbf{P}_1) \mathbf{Y}\|^2$. Let $\mathbf{M} = \mathbf{I} - \mathbf{P}_1 = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T$. This is the projection matrix onto the orthogonal complement of $\text{span}\{\mathbf{1}\}$. This subspace has dimension $n - 1$. \mathbf{M} is often called the "centering matrix". The sum of squares is $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \mathbf{Y}^T \mathbf{M}^T \mathbf{M} \mathbf{Y} = \mathbf{Y}^T \mathbf{M} \mathbf{Y}$ (since \mathbf{M} is symmetric and idempotent). So, $(n - 1) S_n^2 = \mathbf{Y}^T \mathbf{M} \mathbf{Y}$.

If $Y_i \sim N(\mu, \sigma^2)$ independently, then $\mathbf{Y} \sim N(\mu \mathbf{1}, \sigma^2 \mathbf{I})$. The quadratic form $(n - 1) S_n^2 / \sigma^2 = \frac{1}{\sigma^2} \mathbf{Y}^T \mathbf{M} \mathbf{Y}$ involves the projection matrix \mathbf{M} . If $\mu = 0$, then $\frac{1}{\sigma^2} \mathbf{Y}^T \mathbf{M} \mathbf{Y} \sim \chi_{\text{rank}(\mathbf{M})}^2$. Since $\text{rank}(\mathbf{M}) = \text{rank}(\mathbf{I}) - \text{rank}(\mathbf{P}_1) = n - 1$, we get $\frac{(n-1) S_n^2}{\sigma^2} \sim \chi_{n-1}^2$. This result (Cochran's Theorem) holds even if $\mu \neq 0$. The fact that S_n^2 is an unbiased estimator for σ^2 ($E[S_n^2] = \sigma^2$) relies on $E[\mathbf{Y}^T \mathbf{M} \mathbf{Y}] = \text{tr}(\mathbf{M}(\sigma^2 \mathbf{I})) + (\mu \mathbf{1})^T \mathbf{M}(\mu \mathbf{1}) = \sigma^2 \text{tr}(\mathbf{M}) + \mu^2 \mathbf{1}^T \mathbf{M} \mathbf{1}$. Since $\mathbf{M} \mathbf{1} = \mathbf{0}$ and $\text{tr}(\mathbf{M}) = n - 1$, we get $E[(n - 1) S_n^2] = \sigma^2(n - 1)$, proving unbiasedness.

8 Vector and Matrix Calculus

Derivatives involving vectors and matrices are essential in optimization problems (like OLS).

Definition 8.1 (Gradient). For a scalar-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f(x_1, \dots, x_n)$, the **gradient** is the vector of partial derivatives:

$$\nabla f = \frac{\partial f}{\partial \mathbf{x}} = \begin{pmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_n \end{pmatrix}$$

Definition 8.2 (Derivative of Vector w.r.t. Scalar). For a vector-valued function $\mathbf{y} : \mathbb{R} \rightarrow \mathbb{R}^m$, $\mathbf{y}(x) = (y_1(x), \dots, y_m(x))^T$, the derivative with respect to the scalar x is:

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{pmatrix} \partial y_1 / \partial x \\ \vdots \\ \partial y_m / \partial x \end{pmatrix}$$

Definition 8.3 (Jacobian Matrix). For a vector-valued function $\mathbf{y} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), \dots, y_m(\mathbf{x}))^T$ where $\mathbf{x} = (x_1, \dots, x_n)^T$, the derivative of \mathbf{y} with respect to \mathbf{x} is the $m \times n$ **Jacobian matrix**:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} = [\nabla y_1^T, \dots, \nabla y_m^T]^T$$

(Note: Some definitions transpose this matrix. We use the layout where rows correspond to components of \mathbf{y} and columns to components of \mathbf{x} .)

Proposition 8.1 (Common Vector Derivatives). Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^n$ be a constant vector, and $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a constant matrix.

1. Linear function (scalar output): $f(\mathbf{x}) = \mathbf{b}^T \mathbf{x} = \mathbf{x}^T \mathbf{b}$.

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^T \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T \mathbf{b}) = \mathbf{b}$$

2. Quadratic form (scalar output): $f(\mathbf{x}) = \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|^2$.

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T \mathbf{x}) = 2\mathbf{x}$$

3. Linear transformation (vector output): $\mathbf{y}(\mathbf{x}) = \mathbf{A}\mathbf{x}$.

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{A}\mathbf{x}) = \mathbf{A}$$

(This is the Jacobian matrix. The i -th row is $\frac{\partial}{\partial \mathbf{x}}(\sum_j A_{ij}x_j)$, whose k -th element is $\frac{\partial}{\partial x_k}(\sum_j A_{ij}x_j) = A_{ik}$. So the i -th row is the i -th row of \mathbf{A} .)

4. General quadratic form (scalar output): $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{n \times n}$.

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

If \mathbf{A} is symmetric ($\mathbf{A} = \mathbf{A}^T$), then:

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T \mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$$

Derivation of $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T \mathbf{A}\mathbf{x})$. Let $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n A_{ij}x_i x_j$. We find the k -th component of the gradient:

$$\begin{aligned} \frac{\partial f}{\partial x_k} &= \frac{\partial}{\partial x_k} \left(\sum_{i=1}^n \sum_{j=1}^n A_{ij}x_i x_j \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n A_{ij} \frac{\partial}{\partial x_k}(x_i x_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n A_{ij}(\delta_{ik}x_j + x_i\delta_{jk}) \quad (\text{where } \delta_{ab} = 1 \text{ if } a = b, 0 \text{ otherwise}) \\ &= \sum_{j=1}^n A_{kj}x_j + \sum_{i=1}^n A_{ik}x_i \\ &= (\text{k-th row of } \mathbf{A})\mathbf{x} + (\text{k-th column of } \mathbf{A})^T \mathbf{x} \\ &= [\mathbf{A}\mathbf{x}]_k + [\mathbf{A}^T \mathbf{x}]_k = [(\mathbf{A} + \mathbf{A}^T)\mathbf{x}]_k \end{aligned}$$

Thus, the gradient vector is $(\mathbf{A} + \mathbf{A}^T)\mathbf{x}$. If \mathbf{A} is symmetric, this simplifies to $2\mathbf{A}\mathbf{x}$. □