Foundations of Linear Algebra for Statistical Modeling: Change of Basis, Transformations, and Projections

A Guide for the Curious Undergraduate

April 2, 2025

Abstract

Welcome! This document revisits fundamental concepts from linear algebra, viewing them through the lens of statistical modeling and regression analysis. We'll explore how vectors and transformations behave under different coordinate systems (bases), delve into the spectral properties of matrices, understand the geometry of matrix operations, and master the crucial concept of orthogonal projections. Our goal is not just to state definitions and theorems, but to build intuition and appreciate the elegance of these mathematical tools.

1 Vectors, Bases, and Coordinate Representations

At its heart, linear algebra deals with vector spaces. While we often work in the familiar $\mathbb{R}^n$ with the standard basis, understanding how to represent vectors and transformations in *different* bases is key to unlocking deeper insights, particularly in dimensionality reduction and model interpretation.

Lemma 1.1 (Matrix-Vector Product as Linear Combination). *Let $A \in \mathbb{R}^{n\times m}$ be a matrix and $u \in \mathbb{R}^m$ be a vector. The product $v = Au$ is a linear combination of the columns of A, specifically $v = \sum_{j=1}^m u_j A^j$, where A^j is the j-th column of A and u_j is the j-th component of u. Consequently, v belongs to the column space of A, i.e., $v \in colspace(A)$.*

Proof. Looking at the i-th component of v, we have $v_i = \sum_{j=1}^m A_{ij}u_j$. Summing over j, this means the i-th component of v is formed by scaling the i-th component of each column A^j by the corresponding u_j and adding them up: $v_i = \sum_{j=1}^m [A^j]_i u_j$. Considering all components simultaneously, we see $v = \sum_{j=1}^m u_j A^j$. □

This simple lemma is surprisingly powerful. It tells us that multiplying by a matrix A maps vectors from $\mathbb{R}^m$ into the subspace spanned by the columns of A in $\mathbb{R}^n$.

Definition 1.1 (Coordinates Relative to a Basis). Let V be an n-dimensional vector space with an ordered basis $B = \{b_1, \ldots, b_n\}$. Any vector $v \in V$ can be uniquely expressed as a linear combination $v = \sum_{i=1}^n k_i b_i$. The vector $[v]_B = (k_1, \ldots, k_n)^T \in \mathbb{R}^n$ is called the **coordinate vector** of v relative to the basis B.

Remark 1.1. If B is the standard basis $E = \{e_1, \ldots, e_n\}$ for $\mathbb{R}^n$, then $[v]_E = v$. The standard basis provides our default Cartesian coordinate system. Choosing a different basis B is like choosing a new coordinate system, potentially tilted or scaled relative to the standard one. The coordinate vector $[v]_B$ tells us how to "reach" v by taking steps along the directions defined by the basis vectors in B.

Example 1.1. Consider $V = \mathbb{R}^2$ with the basis $B = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$ and the vector $v = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$. We want to find $[v]_B = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}$ such that $k_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + k_2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$. This leads to the system:

$$\begin{aligned} k_1 + k_2 &= 2 \\ k_2 &= 5 \end{aligned}$$

Solving this gives $k_2 = 5$ and $k_1 = -3$. Thus, $[v]_B = \begin{pmatrix} -3 \\ 5 \end{pmatrix}$. Alternatively, note that the system is equivalent to $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$. Solving this matrix equation yields the same coordinate vector.

2 Changing Perspectives: Change-of-Basis Matrices

How do the coordinates of a vector change when we switch from one basis to another? This is where the change-of-basis matrix comes in.

Definition 2.1 (Change-of-Basis Matrix)**.** Let $B = \{b_1, \dots, b_n\}$ and $C = \{c_1, \dots, c_n\}$ be two ordered bases for a vector space V. The **change-of-basis matrix** from B to C, denoted $[I]_C^B$, is the unique matrix such that for any $v \in V$,

$$[v]_C = [I]_C^B [v]_B$$

This matrix transforms coordinate vectors relative to B into coordinate vectors relative to C.

Proposition 2.1. *The columns of the change-of-basis matrix $[I]_C^B$ are the coordinate vectors of the basis vectors of B relative to the basis C. That is, the j-th column of $[I]_C^B$ is $[b_j]_C$. Specifically, if we represent the bases B and C as matrices whose columns are the basis vectors (relative to the standard basis, say), then $[I]_C^B = C^{-1}B$.*

Proof. Let $v \in V$. We can write v in terms of basis B using its coordinates $[v]_B = (k_1, \dots, k_n)^T$: $v = \sum_{j=1}^n k_j b_j$. Applying the coordinate map relative to C, $[v]_C = [\sum_{j=1}^n k_j b_j]_C$. Since the coordinate map is linear, $[v]_C = \sum_{j=1}^n k_j [b_j]_C$. This sum is precisely the matrix-vector product of the matrix whose columns are $[b_j]_C$ with the vector $[v]_B$. Thus, the matrix must be $[I]_C^B$. The relation $[I]_C^B = C^{-1}B$ follows directly from the definition: $[v]_C = C^{-1}v$ and $[v]_B = B^{-1}v$. Substituting these into $[v]_C = [I]_C^B [v]_B$ gives $C^{-1}v = [I]_C^B B^{-1}v$. Since this holds for all v, we must have $C^{-1} = [I]_C^B B^{-1}$, which implies $[I]_C^B = C^{-1}B$. □

Remark 2.1. Some sources define the change-of-basis matrix in the opposite direction (from C to B). We adhere to the definition above.

Proposition 2.2. *The change-of-basis matrix from C to B is the inverse of the change-of-basis matrix from B to C:*

$$[I]_B^C = ([I]_C^B)^{-1}$$

Proof. Starting with $[v]_C = [I]_C^B [v]_B$, multiply both sides by $([I]_C^B)^{-1}$ on the left: $([I]_C^B)^{-1}[v]_C = [v]_B$. By the definition of $[I]_B^C$, we must have $[I]_B^C = ([I]_C^B)^{-1}$. □

Corollary 2.1. *If B is a basis represented by the columns of an invertible matrix B, then $B = [I]_E^B$ is the change-of-basis matrix from B to the standard basis E. Consequently, $B^{-1} = [I]_B^E$ is the change-of-basis matrix from the standard basis E to B.*

3 Representing Linear Transformations

Linear transformations map vectors from one space to another while preserving structure (vector addition and scalar multiplication). Just as vectors have different coordinates depending on the basis, linear transformations have different matrix representations depending on the bases chosen for the domain and codomain.

Definition 3.1 (Matrix Representation of a Linear Transformation)**.** Let $T : V \to W$ be a linear transformation, where $B = \{b_1, \ldots, b_m\}$ is a basis for V and $C = \{c_1, \ldots, c_n\}$ is a basis for W. The **matrix representation** of T with respect to bases B and C, denoted $[T]_C^B$, is the unique $n \times m$ matrix such that for any $v \in V$:

$$[T(v)]_C = [T]_C^B [v]_B$$

This matrix maps the coordinates of a vector v in the basis B to the coordinates of its image $T(v)$ in the basis C.

Remark 3.1. The change-of-basis matrix $[I]_C^B$ is a special case where $V = W$ and T is the identity transformation $I(v) = v$.

Proposition 3.1 (Constructing the Matrix Representation)**.** *The columns of $[T]_C^B$ are the coordinate vectors of the images of the basis vectors of B under T, relative to the basis C. That is, the j-th column of $[T]_C^B$ is $[T(b_j)]_C$.*

Proof. Let $v \in V$ with $[v]_B = (k_1, \ldots, k_m)^T$, so $v = \sum_{j=1}^m k_j b_j$. Since T is linear, $T(v) = T(\sum_{j=1}^m k_j b_j) = \sum_{j=1}^m k_j T(b_j)$. Now, express $T(v)$ in coordinates relative to C: $[T(v)]_C = [\sum_{j=1}^m k_j T(b_j)]_C$. Using linearity of the coordinate map, $[T(v)]_C = \sum_{j=1}^m k_j [T(b_j)]_C$. This is precisely the matrix-vector product of the matrix whose columns are $[T(b_j)]_C$ with the vector $[v]_B$. Thus, $[T]_C^B$ must be this matrix. □

How does the matrix representation of T change if we change the bases?

Theorem 3.1 (Change of Basis for Transformations)**.** *Let $T : V \to V$ be a linear operator, and let B and C be two bases for V. Let $[T]_B$ be the matrix representation of T relative to basis B (i.e., $[T]_B^B$) and $[T]_C$ be the matrix representation relative to basis C. Then:*

$$[T]_C = [I]_C^B [T]_B [I]_B^C$$

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

Proof. We want to show that for any $v \in V$, $[T]_C [v]_C = ([I]_C^B [T]_B [I]_B^C)[v]_C$. Starting with the right side:

$$\begin{aligned}
([I]_C^B [T]_B [I]_B^C)[v]_C &= [I]_C^B [T]_B ([I]_B^C [v]_C) \\
&= [I]_C^B [T]_B [v]_B \quad \text{(since } [I]_B^C \text{ converts } C\text{-coords to } B\text{-coords)} \\
&= [I]_C^B [T(v)]_B \quad \text{(by definition of } [T]_B) \\
&= [T(v)]_C \quad \text{(since } [I]_C^B \text{ converts } B\text{-coords to } C\text{-coords)}
\end{aligned}$$

The left side is $[T]_C [v]_C$, which by definition is also equal to $[T(v)]_C$. Since the equality holds for all $[v]_C$, the matrices must be equal. □

Remark 3.2. Matrices A and B related by $B = P^{-1}AP$ for some invertible matrix P are called **similar matrices**. Theorem 3.1 states that matrices representing the same linear operator T with respect to different bases are similar. The change-of-basis matrix plays the role of P^{-1} (or P, depending on the direction). In the example provided in the notes, $F(x, y) = (5x - y, 2x + y)$. The matrix A representing F in the standard basis E is $A = [F]_E = \begin{pmatrix} 5 & -1 \\ 2 & 1 \end{pmatrix}$. The basis $S = \{u_1, u_2\} = \{(1, 4), (2, 7)\}$. The change-of-basis matrix from S to E is $P = [I]_E^S = \begin{pmatrix} 1 & 2 \\ 4 & 7 \end{pmatrix}$ (its columns are u_1, u_2). The change-of-basis matrix from E to S is $P^{-1} = [I]_S^E = \begin{pmatrix} -7 & 2 \\ 4 & -1 \end{pmatrix}$. The matrix B representing F in the basis S is $B = [F]_S = [I]_S^E [F]_E [I]_E^S = P^{-1}AP$. $B = \begin{pmatrix} -7 & 2 \\ 4 & -1 \end{pmatrix} \begin{pmatrix} 5 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 4 & 7 \end{pmatrix} = \begin{pmatrix} -31 & 9 \\ 18 & -5 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 4 & 7 \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ -2 & 1 \end{pmatrix}$.

Remark 3.3 (Standard Representation)**.** Unless specified otherwise, a given matrix A is assumed to represent a linear transformation $T(v) = Av$ with respect to the standard basis in both the domain and codomain. In this standard representation:

- The Kernel (Null Space) of T corresponds to the null space of A: $\text{Ker}(T) = \text{Ker}(A) = \{v \mid Av = 0\}$.
- The Image (Range) of T corresponds to the column space of A: $\text{Im}(T) = \text{Im}(A) = \text{colspace}(A) = \{Av \mid v \in V\}$.

Finding a basis for $\text{Ker}(T)$ amounts to finding a basis for the solution space of the homogeneous system $Av = 0$. Finding a basis for $\text{Im}(T)$ involves finding a basis for the column space, often done by finding the pivot columns of A or by row-reducing A^T and taking the non-zero rows as the basis vectors.

Theorem 3.2 (Rank-Nullity Theorem for Matrices)**.** *For any matrix $A \in \mathbb{R}^{n \times m}$, let rank$(A) = r = \dim(Im(A))$ be the rank (dimension of the column space/image) and $\dim(Ker(A))$ be the nullity (dimension of the kernel). Then:*

$$\dim(Im(A)) + \dim(Ker(A)) = m \quad (\textit{dimension of the domain})$$

Or, in terms of rank: $r + \dim(Ker(A)) = m$, so $\dim(Ker(A)) = m - r$. (Note: The original text seems to have a typo, stating $\dim(Ker(A)) = n - r$; it should be $m - r$, where m is the number of columns/dimension of the domain).

4 Eigenvalues, Eigenvectors, and Spectral Decomposition

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

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

$$Ax = \lambda x$$

The vector x is called an **eigenvector** corresponding to the eigenvalue λ.

Intuitively, eigenvectors x are special vectors whose direction is unchanged (or flipped if $\lambda < 0$) by the transformation A; they are only stretched or shrunk by a factor of λ.

Definition 4.2 (Characteristic Polynomial)**.** The **characteristic polynomial** of a square matrix $A \in \mathbb{R}^{n \times n}$ is defined as

$$p(\lambda) = \det(A - \lambda I)$$

where I is the identity matrix.

Theorem 4.1. *The eigenvalues of A are precisely the roots of the characteristic polynomial $p(\lambda) = 0$.*

Proof. The equation $Ax = \lambda x$ can be rewritten as $Ax - \lambda Ix = 0$, or $(A - \lambda I)x = 0$. For this equation to have a non-zero solution x (as required for an eigenvector), the matrix $(A - \lambda I)$ must be singular (non-invertible). This occurs exactly when its determinant is zero: $\det(A - \lambda I) = 0$. □

4.1 Spectral Decomposition for Symmetric Matrices

Symmetric matrices ($A = A^T$) have particularly nice properties regarding their eigenvalues and eigenvectors, leading to a powerful decomposition.

Theorem 4.2 (Spectral Theorem for Symmetric Matrices). *If $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, then:*

1. *All eigenvalues of A are real.*

2. *Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

3. *There exists an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of A.*

4. *A can be diagonalized by an orthogonal matrix U. That is, there exists an orthogonal matrix U (whose columns are the orthonormal eigenvectors of A) and a diagonal matrix Λ (whose diagonal entries are the corresponding eigenvalues) such that:*

$$A = U\Lambda U^T$$

This is known as the ***spectral decomposition*** *of A. Equivalently, $U^T AU = \Lambda = diag(\lambda_1, \dots, \lambda_n)$.*

Remark 4.1 (Interpretation). The spectral decomposition $A = U\Lambda U^T$ tells us that the action of a symmetric matrix A can be understood as a sequence of three simpler operations:

1. **Rotation/Reflection (U^T):** Change from the standard basis to the orthonormal eigenbasis (columns of U). Since U is orthogonal, $U^T = U^{-1}$, this is just a rotation or reflection.

2. **Scaling (Λ):** Scale along the axes of the new coordinate system (the eigenvectors). The scaling factor along the j-th eigen-axis is the j-th eigenvalue λ_j.

3. **Rotation/Reflection Back (U):** Change back from the eigenbasis to the standard basis.

This relates to the geometric interpretation discussed later and visualized in the provided figures.

Example 4.1 (Exercise Interpretation from Notes). Let $A = X^T X$ where $X \in \mathbb{R}^{n\times p}$. A is symmetric ($A^T = (X^T X)^T = X^T(X^T)^T = X^T X = A$) and belongs to $\mathbb{R}^{p\times p}$ (assuming n was a typo in the notes). Let its spectral decomposition be $A = U\Lambda U^T$, where U is a $p \times p$ orthogonal matrix of eigenvectors and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ contains the eigenvalues.

- Consider the transformation $T : \mathbb{R}^p \to \mathbb{R}^p$ defined by $T(v) = Av$. The matrix representing T in the standard basis E is simply $A = [T]_E$.

- What is the matrix representing T in the basis $B = U$ (the orthonormal basis of eigenvectors)? Using Theorem 3.1, the matrix is $[T]_U = [I]_U^E [T]_E [I]_E^U$. Here, $[I]_E^U = U$ (transforms from U-basis to E-basis) and $[I]_U^E = U^{-1} = U^T$ (transforms from E-basis to U-basis). So, $[T]_U = U^T AU = U^T(U\Lambda U^T)U = (U^T U)\Lambda(U^T U) = I\Lambda I = \Lambda$. This confirms that in the basis of its eigenvectors, the transformation A acts simply by scaling along those eigenvector directions by the corresponding eigenvalues.

- The transformation $f : \mathbb{R}^p \to \mathbb{R}^p$ given by $f(v) = (\lambda_1 v_1, \ldots, \lambda_p v_p)^T$ represents scaling along the standard coordinate axes. Its matrix in the standard basis is Λ. The fact that A (representing T in basis E) is similar to Λ (representing T in basis U, or f in basis E) via $A = U\Lambda U^T$ highlights that T *is* fundamentally a scaling operation, but performed along the directions defined by U, not necessarily the standard axes.

Remark 4.2 (Geometric Interpretation of $A = U\Sigma V^T$ (SVD))**.** While the notes mention spectral decomposition ($A = U\Lambda U^T$) for symmetric A, the geometric interpretation often uses the Singular Value Decomposition (SVD), $A = U\Sigma V^T$, which applies to *any* matrix $A \in \mathbb{R}^{n\times m}$. Here U ($n \times n$) and V ($m \times m$) are orthogonal matrices, and Σ ($n \times m$) is a diagonal matrix of singular values. The action $w = Av$ can be seen as:

1. **Rotation/Reflection (V^T):** Change basis in the domain using V^T.
2. **Scaling (Σ):** Scale along the new axes (singular values). Also handles change in dimension if $n \neq m$.
3. **Rotation/Reflection (U):** Change basis in the codomain using U.

This "Rotate-Scale-Rotate" sequence is a powerful way to visualize any linear transformation. For symmetric positive semi-definite matrices (like $X^T X$), the SVD and spectral decomposition are closely related.

4.2 Properties Related to Eigenvalues

For any square matrix A that is diagonalizable (similar to a diagonal matrix, which includes all symmetric matrices), its eigenvalues hold information about its determinant and trace.

Proposition 4.1 (Determinant and Trace from Eigenvalues)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a diagonalizable matrix (e.g., symmetric) with eigenvalues $\lambda_1, \ldots, \lambda_n$ (counted with multiplicity). Then:*

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

where tr$(A) = \sum_{i=1}^n A_{ii}$ is the trace of A (the sum of its diagonal elements).

Proof. Since A is diagonalizable, $A = P\Lambda P^{-1}$ for some invertible P and diagonal $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

1. $\det(A) = \det(P\Lambda P^{-1}) = \det(P)\det(\Lambda)\det(P^{-1}) = \det(P)\det(\Lambda)(\det(P))^{-1} = \det(\Lambda)$. The determinant of a diagonal matrix is the product of its diagonal entries, so $\det(\Lambda) = \prod_{i=1}^n \lambda_i$.
2. The trace has the cyclic property $\text{tr}(XYZ) = \text{tr}(ZXY)$. So, $\text{tr}(A) = \text{tr}(P\Lambda P^{-1}) = \text{tr}(P^{-1}P\Lambda) = \text{tr}(I\Lambda) = \text{tr}(\Lambda)$. The trace of a diagonal matrix is the sum of its diagonal entries, so $\text{tr}(\Lambda) = \sum_{i=1}^n \lambda_i$.

□

5 Orthogonal Projection Matrices

Projection matrices are fundamental in regression (think OLS) and many areas of statistics and machine learning. They project vectors onto specific subspaces.

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

Applying an idempotent matrix twice has the same effect as applying it once – a characteristic property of projections.

Definition 5.2 (Orthogonal Projection Matrix)**.** A matrix $P \in \mathbb{R}^{n \times n}$ is an **orthogonal projection matrix** if it is both **symmetric** ($P^T = P$) and **idempotent** ($P^2 = P$). 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 P are either 0 or 1. The eigenvalue 1 has multiplicity equal to the rank of P (i.e., the dimension of the subspace onto which it projects), $r = rank(P)$. The eigenvalue 0 has multiplicity equal to the dimension of the kernel of P, which is $n - r$.*

Proof. Let λ be an eigenvalue with eigenvector $x \neq 0$, so $Px = \lambda x$. Apply P again: $P^2 x = P(\lambda x) = \lambda(Px) = \lambda(\lambda x) = \lambda^2 x$. Since P is idempotent, $P^2 = P$, so $P^2 x = Px = \lambda x$. Thus, we must have $\lambda^2 x = \lambda x$, which implies $(\lambda^2 - \lambda)x = 0$. Since $x \neq 0$, we need $\lambda^2 - \lambda = 0$, or $\lambda(\lambda - 1) = 0$. The only possible eigenvalues are $\lambda = 0$ or $\lambda = 1$. The dimension of the eigenspace for $\lambda = 1$ is the dimension of the subspace $\{x \mid Px = x\}$, which is the image of P. The dimension of the eigenspace for $\lambda = 0$ is the dimension of the subspace $\{x \mid Px = 0\}$, which is the kernel of P. By the Rank-Nullity theorem, these dimensions sum to n. □

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

$$P_X = X(X^T X)^{-1} X^T$$

is the orthogonal projection matrix onto the column space of X, $colspace(X) = Im(X)$.

Proof. We need to verify symmetry, idempotency, and that its image is colspace(X).

1. **Symmetry:** $P_X^T = (X(X^T X)^{-1} X^T)^T = (X^T)^T((X^T X)^{-1})^T X^T = X((X^T X)^T)^{-1} X^T = X(X^T(X^T)^T)^{-1} X^T = X(X^T X)^{-1} X^T = P_X$. (We used $(AB)^T = B^T A^T$, $(A^{-1})^T = (A^T)^{-1}$, and $(A^T)^T = A$. Note $(X^T X)$ is symmetric). So P_X is symmetric.

2. **Idempotency:** $P_X^2 = (X(X^T X)^{-1} X^T)(X(X^T X)^{-1} X^T) = X(X^T X)^{-1}(X^T X)(X^T X)^{-1} X^T = XI(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = P_X$. So P_X is idempotent.

3. **Image:** For any $v \in \mathbb{R}^n$, $P_X v = X[(X^T X)^{-1} X^T v]$. Let $w = (X^T X)^{-1} X^T v$. Then $w \in \mathbb{R}^p$, and $P_X v = Xw$, which is a linear combination of the columns of X. Thus, $P_X v \in$ colspace(X). (It can be further shown that if $v \in$ colspace(X), then $P_X v = v$).

□

5.1 Key Properties of Projection Matrices

Let $P_X = X(X^T X)^{-1} X^T$ be the projection onto colspace(X) (assuming X has full column rank).

1. P_X is symmetric.

2. P_X is idempotent ($P_X^2 = P_X$).

3. $P_X X = X$. (Projection onto colspace(X) leaves vectors already in colspace(X) unchanged. Since columns of X are in colspace(X), they are unchanged.) $P_X X = X(X^T X)^{-1} X^T X = XI = X$.

4. $X^T(I - P_X) = 0$. (The residual vector $v - P_X v$ is orthogonal to the columns of X). $X^T(I - P_X) = X^T - X^T P_X = X^T - X^T X(X^T X)^{-1} X^T = X^T - (X^T X)(X^T X)^{-1} X^T = X^T - IX^T = X^T - X^T = 0$.

5. $P_X v \in \text{Im}(X)$ for all $v \in \mathbb{R}^n$.

6. If X is $n \times n$ and invertible, then colspace$(X) = \mathbb{R}^n$, and $P_X = X(X^TX)^{-1}X^T = XX^{-1}(X^T)^{-1}X^T = II = I$. Projection onto the whole space is the identity map.

7. $(I - P_X)v \in \text{Im}(X)^\perp$ for all $v \in \mathbb{R}^n$. (The residual vector is orthogonal to the projection space). Let $w \in \text{Im}(X)$, so $w = Xz$ for some z. Then $w^T(I - P_X)v = (Xz)^T(I - P_X)v = z^TX^T(I - P_X)v$. From property 4, $X^T(I - P_X) = 0$, so the dot product is 0.

8. If $w \in \text{Im}(X)$, then $P_X w = w$.

9. If $w \in \text{Im}(X)^\perp$, then $P_X w = 0$. (Vectors orthogonal to the space project to the origin). $P_X w = X(X^TX)^{-1}X^Tw$. Since $w \in \text{Im}(X)^\perp$, $X^Tw = 0$. Thus $P_X w = 0$.

10. P_X depends only on the subspace $\text{Im}(X)$, not the specific basis X. If $\text{Im}(Z) = \text{Im}(X)$, then $P_Z = P_X$.

11. If $L \subseteq M$ are subspaces, then $P_M P_L = P_L P_M = P_L$. (Projecting onto L and then onto the larger space M is the same as just projecting onto L. Projecting onto M first, then L, also results in projection onto L).

Definition 5.3 (Orthogonal Complement). Let V be an inner product space and $U \subseteq V$ be a subspace. The **orthogonal complement** of U, denoted $U^\perp$, is the set of all vectors in V that are orthogonal to every vector in U:

$$U^\perp = \{v \in V \mid u^Tv = 0 \text{ for all } u \in U\}$$

$U^\perp$ is also a subspace of V.

Theorem 5.1 (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 $U^\perp$:*

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

This means every vector $v \in V$ can be uniquely written as $v = u + w$, where $u \in U$ and $w \in U^\perp$. (Here, $u = P_U v$ and $w = v - P_U v = (I - P_U)v$).

Proposition 5.3 (Projection onto Orthogonal Complement). *Let P_X be the orthogonal projection onto $M = Im(X)$. Then $I - P_X$ is the orthogonal projection onto the orthogonal complement $M^\perp = Im(X)^\perp$. That is, $P_{M^\perp} = I - P_M$.*

Proof. Let $Q = I - P_X$.

- Symmetry: $Q^T = (I - P_X)^T = I^T - P_X^T = I - P_X = Q$ (since P_X is symmetric).
- Idempotency: $Q^2 = (I - P_X)(I - P_X) = I - P_X - P_X + P_X^2 = I - P_X - P_X + P_X = I - P_X = Q$ (since P_X is idempotent).

So Q is an orthogonal projection matrix. Its image is $\{(I - P_X)v \mid v \in \mathbb{R}^n\}$, which by property 7 above is precisely $\text{Im}(X)^\perp$. □

Proposition 5.4 (Projection onto Intersection). *If $L \subseteq M$ are subspaces, then $P_M - P_L$ is the orthogonal projection matrix onto $M \cap L^\perp$ (the part of M that is orthogonal to L).*

Proposition 5.5 (Fundamental Subspaces Relationship). *For any matrix $A \in \mathbb{R}^{n \times m}$, the column space of its transpose is the orthogonal complement of its kernel:*

$$Im(A^T) = (Ker(A))^\perp$$

Remark 5.1 (Diagonalizability)**.** A matrix A might fail to be diagonalizable if it has repeated eigenvalues and not enough linearly independent eigenvectors. A formal condition involves minimal polynomials: A is not diagonalizable if for some eigenvalue λ_i, the smallest $k \geq 2$ such that $(A - \lambda_i I)^k = 0$ requires $k \geq 2$. However, symmetric matrices are *always* diagonalizable (by the Spectral Theorem 4.2).

Remark 5.2 (OLS Connection)**.** The matrix $Q = I - P_X$ projects onto the orthogonal complement of colspace(X). Its spectral decomposition involves eigenvalues 0 and 1. If $P_X = U\Lambda U^T$ with $\Lambda = \text{diag}(\underbrace{1, \dots, 1}_{p}, \underbrace{0, \dots, 0}_{n-p})$, then $Q = I - U\Lambda U^T = UIU^T - U\Lambda U^T = U(I - \Lambda)U^T$. The matrix $I - \Lambda$ is diagonal with $\underbrace{0, \dots, 0}_{p}, \underbrace{1, \dots, 1}_{n-p}$ on the diagonal. In Ordinary Least Squares (OLS), we minimize $\|Y - X\beta\|^2$. The solution is $\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY$. The fitted values are $\hat{Y} = X\hat{\beta}_{OLS} = X(X^TX)^{-1}X^TY = P_XY$. The residuals are $e = Y - \hat{Y} = Y - P_XY = (I - P_X)Y = QY$. The minimization finds the vector $X\beta$ in colspace(X) closest to Y, which is precisely the orthogonal projection P_XY. The squared norm of the residuals is $\|e\|^2 = \|(I - P_X)Y\|^2$. Since $I - P_X$ projects onto a subspace of dimension $n - p$ (where $p = \text{rank}(X)$), increasing the dimension p (adding predictors) generally reduces the dimension of the orthogonal complement, thus potentially reducing the residual norm $\|(I - P_X)Y\|^2$. (It cannot increase it).

6 Vector and Matrix Derivatives

Calculus extends naturally to functions involving vectors and matrices, which is crucial for optimization problems like finding OLS estimators.

Definition 6.1 (Gradient)**.** For a scalar-valued function $f(x_1, \dots, x_n)$ of n variables, its **gradient** (with respect to the vector $x = (x_1, \dots, x_n)^T$) is the vector of partial derivatives:

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

Definition 6.2 (Derivative of Vector w.r.t. Scalar)**.** If $y = (y_1, \dots, y_m)^T$ is a vector-valued function of a scalar x, its derivative is:

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

Definition 6.3 (Jacobian Matrix)**.** If $y = (y_1, \dots, y_m)^T$ is a vector-valued function of a vector $x = (x_1, \dots, x_n)^T$, its derivative (the **Jacobian matrix**) is the $m \times n$ matrix of partial derivatives:

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

The entry (i, j) is $\partial y_i / \partial x_j$. (Note: Some conventions use the transpose of this matrix).

6.1 Useful Derivative Identities

Let $x \in \mathbb{R}^n$ be a vector variable, $b \in \mathbb{R}^n$ a constant vector, $A \in \mathbb{R}^{m \times n}$ a constant matrix, and $M \in \mathbb{R}^{n \times n}$ a constant square matrix.

1. **Linear function:** $f(x) = b^T x = x^T b$. Then $\frac{\partial f}{\partial x} = b$.

2. **Quadratic form (simple):** $f(x) = x^T x$. Then $\frac{\partial f}{\partial x} = 2x$.

3. **Matrix-vector product:** $y(x) = Ax$. Then $\frac{\partial y}{\partial x} = A$ (using the Jacobian definition above).

4. **General quadratic form:** $f(x) = x^T Mx$. Then $\frac{\partial f}{\partial x} = (M + M^T)x$.

   Sketch. $f(x) = \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i x_j$. $\frac{\partial f}{\partial x_k} = \frac{\partial}{\partial x_k}(\sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i x_j)$. The terms where x_k appears are when $i = k$ or $j = k$. $\frac{\partial f}{\partial x_k} = \sum_{j=1}^n M_{kj} x_j + \sum_{i=1}^n M_{ik} x_i$. This is the k-th component of Mx plus the k-th component of $M^T x$, which is the k-th component of $(M + M^T)x$. □

5. **Symmetric quadratic form:** If M is symmetric ($M = M^T$), then $f(x) = x^T Mx$ gives $\frac{\partial f}{\partial x} = (M + M)x = 2Mx$.

These identities are essential for deriving estimators in statistics, such as finding the β that minimizes the sum of squared errors $S(\beta) = (Y - X\beta)^T (Y - X\beta)$ in OLS regression.

7 Further Explorations (Exercises from Notes)

Let's briefly touch upon some additional concepts and exercises mentioned.

Example 7.1 (Rank-1 Projection)**.** Let $v \in \mathbb{R}^n$, $v \neq 0$. Show that $P = \frac{vv^T}{\|v\|^2}$ is an orthogonal projection matrix.

- Symmetry: $P^T = \frac{(vv^T)^T}{\|v\|^2} = \frac{(v^T)^T v^T}{\|v\|^2} = \frac{vv^T}{\|v\|^2} = P$.
- Idempotency: $P^2 = (\frac{vv^T}{\|v\|^2})(\frac{vv^T}{\|v\|^2}) = \frac{v(v^T v)v^T}{(\|v\|^2)^2} = \frac{v(\|v\|^2)v^T}{(\|v\|^2)^2} = \frac{\|v\|^2 vv^T}{(\|v\|^2)^2} = \frac{vv^T}{\|v\|^2} = P$.

Thus, P is an orthogonal projection matrix. Its image is the span of v, which is a 1-dimensional subspace. Therefore, rank$(P) = 1$. This matrix projects any vector onto the line spanned by v.

Remark 7.1 (Unbiased Sample Variance and Chi-Squared)**.** The notes mention showing that $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ is an unbiased estimator for σ^2 when $Y_i \sim (\mu, \sigma^2)$ are i.i.d.. This involves expectation calculations. Furthermore, if $Y_i \sim N(\mu, \sigma^2)$, then $(n-1)S_n^2/\sigma^2 \sim \chi_{n-1}^2$ (a chi-squared distribution with $n-1$ degrees of freedom). This result is foundational in statistical inference (confidence intervals, hypothesis tests for variance). The proof often involves recognizing $\sum(Y_i - \bar{Y})^2$ as a quadratic form $Y^T QY$ where Q is related to the projection matrix $I - P_{\mathbf{1}}$ ($\mathbf{1}$ being a vector of ones) and applying Cochran's theorem.

We hope this review, blending rigor with intuition, illuminates the connections between abstract linear algebra and its practical applications in analyzing data and building statistical models. Keep exploring!