2

Linear Spaces and Transformations

2.1 LINEAR TRANSFORMATIONS

Let V and W be two vector spaces. A mapping

$$L : V \to W$$

is said to be *linear* if

- $L(\mathbf{u}+\mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v})$
- $L(c\mathbf{u}) = cL(\mathbf{u})$

for any vectors $\mathbf{u}, \mathbf{v} \in V$ and scalar $c \in \mathbb{R}$.

Example 2.1. Let A be an $m \times n$ matrix and define L_A

$$\begin{array}{rcl} L_A : \mathbb{R}^n & \to & \mathbb{R}^m \\ L_A(\mathbf{u}) & = & A\mathbf{u} \end{array}$$

Clearly L_A is a linear mapping as a consequence of the linearity of matrix multiplication.

Example 2.2. Differentiation, represented by the Jacobian matrix, is a linear mapping. See Section 4.6.3 in Chapter 3.1 for more details.

Example 2.3. Let $V = \mathbb{R}^3$ and $W = \mathbb{R}^2$. The mapping

$$\begin{array}{rcl} T : \mathbb{R}^3 & \to & \mathbb{R}^2 \\ T(x,y,z) & = & (x,y) \end{array}$$

is also linear.

While it is not surprising that matrix multiplication is a linear mapping, it is notable that *every* linear transformation between finite dimensional vector spaces may be represented as multiplication of a vector by an appropriate matrix. This representation is achieved by the introduction of a coordinate system, or basis for the space.

For example, the n vectors

$$\mathbf{e}^{(1)} = (1\,0\,\cdots\,0)^T$$

$$\mathbf{e}^{(2)} = (0\,1\,\cdots\,0)^T$$

$$\mathbf{e}^{(n)} = (0\,0\,\cdots\,1)^T$$

form a basis for $\mathbb{R}^n$ known as the *standard basis*. Thus any $\mathbf{u} \in \mathbb{R}^n$ can be written

$$\mathbf{u} = \alpha_1\mathbf{e}^{(1)} + \alpha_2\mathbf{e}^{(2)} + \cdots + \alpha_n\mathbf{e}^{(n)}.$$

The n-tuple $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ determines the *coordinates* of the point $\mathbf{u}$ w.r.t. to the standard basis.

We digress for a moment to emphasize the dependence of the coordinates of $\mathbf{u}$ on the choice of basis. For example, give another basis $\mathcal{B}$ for $\mathbb{R}^n$ consisting of the vectors $\{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\}$ we may represent $\mathbf{u}$ as

$$\mathbf{u} = x_1\mathbf{v}^{(1)} + \cdots + x_n\mathbf{v}^{(n)}.$$

The n-tuple $(x_1, \dots, x_n)$ determine the coordinates of the point $\mathbf{u}$ w.r.t. to the new basis $\mathcal{B}$. More on this in the following section.

Now that the vector space is equipped with a basis we may make the connection between linear transformations and matrices.

Proposition 2.1. *Every linear mapping can be written as matrix multiplication.*

Proof. Consider the linear mapping

$$L : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Let $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \dots, \mathbf{e}^{(n)}$ be the standard basis for $\mathbb{R}^n$. Furthermore, let $L(\mathbf{e}^{(j)}) = \mathbf{a}_j$ where $\mathbf{a}_j$ is a column vector in $\mathbb{R}^m$. Specifically, $L(\mathbf{e}^{(i)}) = (a_{1i}\, a_{2i}\, \dots\, a_{mi})^T$. Now let $\mathbf{u}$ be an arbitrary element of $\mathbb{R}^n$, i.e., $\mathbf{u} = \alpha_1\mathbf{e}^{(1)} + \cdots + \alpha_n\mathbf{e}^{(n)}$. Thus we have

$$\begin{aligned} L(\mathbf{u}) &= L(\alpha_1\mathbf{e}^{(1)} + \cdots + \alpha_n\mathbf{e}^{(n)}) \\ &= \alpha_1 L(\mathbf{e}^{(1)}) + \cdots + \alpha_n L(\mathbf{e}^{(n)}) \\ &= \alpha_1\mathbf{a}_1 + \cdots + \alpha_n\mathbf{a}_n \\ &= A\alpha \end{aligned}$$

where $(A)_{ij} = a_{ij}$. □

Example 2.4. The matrix which corresponds to the linear operator of Example 2.3 is given by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

2.2 CHANGE OF BASIS

A central issue in studying patterns is determining and utilizing the *correct* basis for a given set of data. Later we argue that empirical bases tend to be more efficient for representing specific data sets.

Motivated by this we now develop the basic mechanics of changing coordinate systems. To start, let $\{\mathbf{v}^{(i)}\}_{i=1}^n$ and $\{\mathbf{w}^{(i)}\}_{i=1}^n$ both be bases for $\mathbb{R}^n$, called $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively. Let $\mathbf{u}$ be an arbitrary element of $\mathbb{R}^n$. Thus in terms of the basis $\mathcal{B}_1$ we write

$$\mathbf{u} = x_1\mathbf{v}^{(1)} + x_2\mathbf{v}^{(2)} + \cdots + x_n\mathbf{v}^{(n)}$$

and in terms of $\mathcal{B}_2$ we write

$$\mathbf{u} = y_1\mathbf{w}^{(1)} + y_2\mathbf{w}^{(2)} + \cdots + y_n\mathbf{w}^{(n)}$$

giving the representation, or coordinates

$$\mathbf{v}_{\mathcal{B}_1} = (x_1 \ \ldots \ x_n)^T$$

w.r.t. $\mathcal{B}_1$ and coordinates

$$\mathbf{u}_{\mathcal{B}_2} = (y_1 \ \ldots \ y_n)^T$$

w.r.t. $\mathcal{B}_2$. Generally, the coordinate system which is in use is clear from the context and no specific reference is made to it.

By assumption, the $\{\mathbf{v}^{(i)}\}$ form a basis for $\mathbb{R}^n$, and any element in $\mathbb{R}^n$ can be expressed in terms of them. Thus, we may write

$$\mathbf{w}^{(i)} = \sum_{j=1}^{n} q_{ij}\mathbf{v}^{(j)}$$

which leads to

$$\begin{aligned}\mathbf{u} &= y_1(\sum_{j=1}^{n} q_{1j}\mathbf{v}^{(j)}) + \cdots + y_n(\sum_{j=1}^{n} q_{nj}\mathbf{v}^{(j)}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} y_i q_{ij}\mathbf{v}^{(j)} = \sum_{j=1}^{n} x_j\mathbf{v}^{(j)}\end{aligned}$$

where $x_j = \sum_{i=1}^{n} q_{ij}y_i$ which is equivalent to

$$\mathbf{x} = Q^T\mathbf{y}$$

which we could equivalently write as $\mathbf{u}_{\mathcal{B}_1} = Q^T\mathbf{u}_{\mathcal{B}_2}$.

Alternatively, we can write $\mathbf{v}^{(i)} = \sum_{j=1}^{n} p_{ij}\mathbf{w}^{(j)}$ which leads to the relationship

$$\mathbf{y} = \mathbf{P}^T\mathbf{x} \tag{2.1}$$

from which it follows that $(\mathbf{P}^T)^{-1} = \mathbf{Q}^T$, i.e., the coordinate transformation is invertible.

Example 2.5. Given the basis vectors defining $\mathcal{B}_1$ to be

$$\mathbf{v}^{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{v}^{(2)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and that the basis vectors defining $\mathcal{B}_2$ are

$$\mathbf{w}^{(1)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{w}^{(2)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

find $\mathbf{u}_{\mathcal{B}_2}$ given

$$\mathbf{u}_{\mathcal{B}_1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

We know that $\mathbf{u}_{\mathcal{B}_2} = P^T \mathbf{u}_{\mathcal{B}_1}$ so we must first compute P. By definition,

$$\mathbf{v}^{(i)} = p_{i1}\mathbf{w}^{(1)} + p_{i2}\mathbf{w}^{(2)}$$

so for $i = 1$ we have

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} p_{11} \\ p_{12} \end{pmatrix}$$

so

$$\begin{pmatrix} p_{11} \\ p_{12} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

From $i = 2$ it follows that

$$\begin{pmatrix} p_{21} \\ p_{22} \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

so

$$P^T = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$$

Given $\mathbf{u}_{\mathcal{B}_2} = P^T \mathbf{u}_{\mathcal{B}_1}$ we have

$$\mathbf{u}_{\mathcal{B}_2} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$$

The next question we address is how does the matrix A change with a change of basis. Assuming the matrix A is defined w.r.t. $\mathcal{B}_1$, what is the representation A' for this matrix w.r.t. the new basis $\mathcal{B}_2$.

Let's first look at the action of the linear transformation A w.r.t. the basis $\mathcal{B}_1$, i.e.,

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

Both $\mathbf{z}$ and $\mathbf{x}$ are coordinates w.r.t. $\mathcal{B}_1$. If we let $\mathbf{z}'$ and $\mathbf{x}'$ be the coordinates of $\mathbf{z}$ and $\mathbf{x}$ w.r.t. $\mathcal{B}_2$, respectively, then there exists a matrix M such that

$$\mathbf{z} = M\mathbf{z}' \quad \text{and} \quad \mathbf{x} = M\mathbf{x}'$$

So $z = Ax$ may be written $M\mathbf{z}' = AM\mathbf{x}'$ or

$$\mathbf{z}' = M^{-1}AM\mathbf{x}'$$

from which we conclude that

$$A' = M^{-1}AM$$

Thus $\mathbf{z}$ is the result of applying A to $\mathbf{x}$ in the first coordinate system and $\mathbf{z}'$ is the result of applying A' in the second coordinate system. In this case A and A' are said to be *similar* matrices.

Example 2.6. Given the bases as defined in Example 2.5 and that the mapping

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

is defined w.r.t. $\mathcal{B}_1$. What is the corresponding transformation w.r.t. $\mathcal{B}_2$? We saw previously that

$$M^{-1} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$$

so

$$A' = M^{-1}AM = \begin{pmatrix} -1 & 6 \\ 0 & 3 \end{pmatrix}$$

The question arises naturally, is there a coordinate system such that the action of a matrix is especially simple? The answer is yes for a large class of well-defined matrices. Suppose that the $n \times n$ matrix has n linearly independent eigenvectors. If these eigenvectors are chosen to be the columns of the transformation matrix M then the new matrix is diagonal, i.e.,

$$M^{-1}AM = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

where the λ_i are the eigenvalues associated with the independent eigenvectors. Note that if $A = A^T$, i.e., A is a symmetric matrix, then A is always diagonalizable. These ideas will be discussed further in Section 2.8.

2.3 OPERATIONS ON SUBSPACES

Given data sets lie initially within large vector spaces, it is important to be able to decompose such spaces into smaller ones. In this section we further develop our tools for decomposing patterns into especially useful subspaces. One of the main ideas to be developed is that of the projection matrix, but first, we examine the general problem of decomposing a vector space into the sum of independent subspaces.

Definition 2.1. *A subspace W of a vector space V is a subset of vectors such that*

- *if $\mathbf{w}, \mathbf{w}' \in W$ and $a, b \in \mathbb{R}$ then $a\mathbf{w} + b\mathbf{w}' \in W$. In this case we say W is closed.*
- *$\mathbf{0} \in W$, i.e., every subspace must contain the zero vector.*

Proposition 2.2. *Define the set of vectors W*

$$W = \{\mathbf{w} \,:\, \mathbf{w} = \sum_i \alpha_i \mathbf{v}^{(i)}\}$$

The set W is a subspace. W is said to be spanned by the set of vectors $\{\mathbf{v}^{(i)}\}$.

Proof. Let $\mathbf{w}, \mathbf{w}' \in W$ so $\mathbf{w} = \sum_i c_i \mathbf{v}^{(i)}$ and $\mathbf{w}' = \sum_i c_i' \mathbf{v}^{(i)}$. It follows that

$$\mathbf{w} + \mathbf{w}' = \sum_i (c_i + c_i')\mathbf{v}^{(i)} \in W$$

$$a\mathbf{w} = \sum (ac_i)\mathbf{v}^{(i)} \in W$$

and lastly

$$\mathbf{0} = \sum_i 0\mathbf{v}^{(i)} \in W.$$

□

2.3.1 Intersection of Subspaces

Proposition 2.3. *If W_1 and W_2 are both subspaces, then so is their intersection $W_1 \cap W_2$.*

Proof. Let $\mathbf{x}, \mathbf{y} \in W_1 \cap W_2$, i.e., $\mathbf{x} \in W_1, W_2$ and $\mathbf{y} \in W_1, W_2 \Rightarrow a\mathbf{x} + b\mathbf{y} \in W_1$ and $a\mathbf{x} + b\mathbf{y} \in W_2$, in other words $a\mathbf{x} + b\mathbf{y} \in W_1 \cap W_2$. □

2.3.2 Addition of Subspaces

Let W_1 and W_2 be subspaces of a vector space V. The sum of these spaces is defined to be the result of taking all possible combinations of the elements of these two spaces.

Definition 2.2. *The sum of the vector subspaces W_1 and W_2 is written $W = W_1 + W_2$ and is defined to be the set*

$$W_1 + W_2 = \{\mathbf{w}_1 + \mathbf{w}_2 : \mathbf{w}_1 \in W_1, \mathbf{w}_2 \in W_2\}$$

The sum of three or more subspaces is defined analogously.

Proposition 2.4. *The sum of two subspaces is a subspace.*

Proof. Let $\mathbf{x}, \mathbf{y} \in W$, i.e., $\mathbf{x} = \mathbf{w}_1 + \mathbf{w}_2$ and $\mathbf{y} = \mathbf{w}_1' + \mathbf{w}_2'$ where $\mathbf{w}_i, \mathbf{w}_i' \in W_i$. Then we have $\alpha\mathbf{x} + \beta\mathbf{y} = \alpha(\mathbf{w}_1 + \mathbf{w}_1') + \beta(\mathbf{w}_2 + \mathbf{w}_2') = \alpha\mathbf{w}_1'' + \beta\mathbf{w}_2'' \in W_1 + W_2$. □.

The fact that the addition of two subspaces is a subspace provides us with a nice way to decompose a vector, i.e., if $\mathbf{x} \in W$ and $W = W_1 + W_2$ we can always write $\mathbf{x} = \mathbf{w}_1 + \mathbf{w}_2$ where $\mathbf{w}_i \in W_i$. After a little bit of experimenting with this decomposition it is clear that it is not unique. This ambiguity will generally be undesirable but can be avoided by restricting the relationship between W_1 and W_2 as described below.

2.3.3 Independence of Subspaces

To make the decomposition of a vector unique we require that the subspaces be independent.

Definition 2.3. *The subspaces W_1 and W_2 of V are independent if*

$$\mathbf{w}_1 + \mathbf{w}_2 = \mathbf{0}$$

implies

$$\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{0}$$

where $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$.

Independence ensures that the decomposition of V into subspaces is unique.

Proposition 2.5. *If W_1, W_2 are independent subspaces and $V = W_1 + W_2$, $\mathbf{w}_1 \in W_1, \mathbf{w}_2 \in W_2$, then the decomposition of $\mathbf{x} \in V$ given by*

$$\mathbf{x} = \mathbf{w}_1 + \mathbf{w}_2$$

is unique.

Proof. Let $\mathbf{x} = \mathbf{w}_1' + \mathbf{w}_2'$ with each $\mathbf{w}_i' \in W_i$. Then

$$\mathbf{x} - \mathbf{x} = \mathbf{0} = (\mathbf{w}_1 - \mathbf{w}_1') + (\mathbf{w}_2 - \mathbf{w}_2').$$

Since W_1 and W_2 are independent we conclude that $\mathbf{w}_i - \mathbf{w}_i' = \mathbf{0}$ or

$$\mathbf{w}_1 = \mathbf{w}_1' \quad \text{and} \quad \mathbf{w}_2 = \mathbf{w}_2'$$

□

Proposition 2.6. *If W_1, W_2 are independent subspaces then*

$$W_1 \cap W_2 = \{\mathbf{0}\}.$$

Proof. Let $\mathbf{w} \in W_1 \cap W_2$. This implies that $\mathbf{w} \in W_1$ and $\mathbf{w} \in W_2$. Since W_2 is a subspace $-\mathbf{w} \in W_2$. Hence

$$\mathbf{w} + (-\mathbf{w}) = \mathbf{0}.$$

Since W_1 and W_2 are independent $\mathbf{w} = -\mathbf{w} = \mathbf{0}$.

Note that the converse is also true. See Problem 2.10.

2.3.4 Direct Sum Decompositions

From above we have that the independence of subspaces and the statement $W_1 \cap W_2 = \{\mathbf{0}\}$ are equivalent. If either (equivalent) properties hold the decomposition is unique and we distinguish the decomposition from the mere addition of subspaces by writing

$$W = W_1 \oplus W_2$$

as the *direct sum decomposition* of W.

These ideas extend directly to the case of more than two subspaces. We cite the following important lemma from [25], p209.

Lemma 2.1. *Let V be a finite dimensional vector space. Let $W_1, \ldots W_k$ be subspaces of V such that $W = W_1 + \cdots + W_k$. The following are equivalent:*

- *$W_1, \ldots, W_k$ are independent.*
- *For each j, $2 \leq j \leq k$,*

$$W_j \cap \{W_1 + \cdots + W_{j-1}\} = \{\mathbf{0}\}$$

- *If $\mathcal{B}_i$ is a basis for W_i then the collection of bases $\{\mathcal{B}_1, \ldots, \mathcal{B}_k\}$ is a basis for W.*

Proof. See [25].

Furthermore, if any (and therefore all) of the above hold then the subspaces W_i form a direct sum decomposition of W which we write

$$V = W_1 \oplus W_2 \oplus \cdots \oplus W_k.$$

2.3.5 Orthogonal Direct Sum Decompositions.

A special but important instance of independent subspaces is orthogonal subspaces.

Definition 2.4. *A vector $\mathbf{v} \in V$ is said to be orthogonal to a subspace $W \subseteq V$ if $\mathbf{v}$ is orthogonal to every $\mathbf{w} \in W$. Two subspaces W_1 and W_2 are said to be orthogonal subspaces if for every $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$ the inner product satisfies $(\mathbf{w}_1, \mathbf{w}_2) = 0$.*

Given a subspace W of the vector space V, the space of all vectors orthogonal to W in V is called the *orthogonal complement* of W written $W^{\perp}$.

Example 2.7. Let $V = \mathbb{R}^3$. Then the x-axis and y-axis are orthogonal subspaces of $\mathbb{R}^3$. Also, the orthogonal complement of the xy-plane is the $z-$axis.

An important special case of the direct sum decomposition occurs when the subspaces are orthogonal. In this situation we distinguish the direct sum notation by writing $\dot{\oplus}$.

Example 2.8. Let V be an n-dimensional vector space with o.n. basis vectors $(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)})$. If $W_i = \mathrm{Span}(\mathbf{v}^{(i)})$ then we can write

$$V = W_1 \dot{\oplus} W_2 \dot{\oplus} \cdots \dot{\oplus} W_n$$

2.4 IMPORTANT SUBSPACES

In this section we describe the basic subspaces which will be of use in what follows. It is implicit, unless otherwise stated, that A is an $m \times n$ matrix.

Definition 2.5. *The range of A, denoted $\mathcal{R}(A)$, is the set of all vectors $\mathbf{v}$ such that $\mathbf{v} = A\mathbf{x}$ i.e.,*

$$\mathcal{R}(A) = \{\mathbf{v} \in \mathbb{R}^m \ : \ \mathbf{v} = A\mathbf{x} \textit{ for some } \mathbf{x} \in \mathbb{R}^n\}$$

The expression $\mathbf{v} = A\mathbf{x}$ may be rewritten

$$\begin{aligned}\mathbf{v} &= [\mathbf{a}^{(1)}|\mathbf{a}^{(2)}|\cdots|\mathbf{a}^{(n)}]\mathbf{x} \\ &= x_1\mathbf{a}^{(1)} + x_2\mathbf{a}^{(2)} + \cdots + x_n\mathbf{a}^{(n)}\end{aligned}$$

This expression reveals the fact that $\mathbf{v}$ lies in the span of the columns of A, i.e.,

$$\mathbf{v} \in \mathrm{Span}\{\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)}\}.$$

Hence the range of A, $\mathcal{R}(A)$, is also referred to as the *column space* of A.

Definition 2.6. *The null space of A, denoted $\mathcal{N}(A)$, is the set of all vectors $\mathbf{y}$ such that $A\mathbf{y} = \mathbf{0}$, i.e.,*

$$\mathcal{N}(A) = \{\mathbf{y} \in \mathbb{R}^n \ : \ A\mathbf{y} = \mathbf{0}\}$$

Definition 2.7. *The row space of A, denoted $\mathcal{R}(A^T)$, is the set of all vectors $\mathbf{x}$ such that $\mathbf{x} = A^T\mathbf{v}$, i.e.,*

$$\mathcal{R}(A^T) = \{\mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{x} = A^T\mathbf{v} \textit{ for some } \mathbf{v} \in \mathbb{R}^m\}$$

Definition 2.8. *The left null space of A, $\mathcal{N}(A^T)$ is the set of all vectors $\mathbf{v}$ such that $A^T\mathbf{v} = \mathbf{0}$, i.e.,*

$$\mathcal{N}(A^T) = \{\mathbf{v} \in \mathbb{R}^m \ : \ A^T\mathbf{v} = \mathbf{0}\}$$

Example 2.9. Find the range $\mathcal{R}(A)$ and null space $\mathcal{N}(A)$ of the matrix

$$A = \mathbf{u}\mathbf{v}^T$$

To determine the range, rewrite the matrix

$$A = [v_1\mathbf{u}|\cdots|v_n\mathbf{u}]$$

from which it is apparent that

$$\mathcal{R}(A) = \{\alpha\mathbf{u} : \alpha \in \mathbb{R}\}$$

Let $\mathbf{x}$ be an element of the null space, i.e.,

$$\mathbf{u}(\mathbf{v}^T\mathbf{x}) = \mathbf{0}$$

From the manner in which this is written we see that, since $\mathbf{u} \neq \mathbf{0}$,

$$\mathcal{N}(A) = \{\mathbf{x} : \mathbf{v}^T\mathbf{x} = 0\}$$

Proposition 2.7. *For any $m \times n$ matrix*

$$\mathcal{N}(A) \perp \mathcal{R}(A^T)$$

i.e., they are orthogonal subspaces of $\mathbb{R}^n$ and

$$\mathcal{N}(A^T) \perp \mathcal{R}(A)$$

i.e., they are orthogonal subspaces of $\mathbb{R}^m$.

The range, or column space of an $m \times n$ matrix A determines a subspace of $\mathbb{R}^m$. The number of independent vectors in this subspace, i.e., its dimension, is a very special and useful quantity for a matrix known as its *rank*.

Definition 2.9. *The column rank (row rank) of a matrix is defined as the number of independent columns (rows) in the column space $\mathcal{R}(A)$ (row space $\mathcal{R}(A^T)$).*

Proposition 2.8. *The row rank is equal to the column rank. In summary,*

$$r = \dim \mathcal{R}(A) = \dim \mathcal{R}(A^T)$$

Example 2.10. The matrix $A = \mathbf{u}\mathbf{v}^T$ has rank $r = 1$.

The following propositions in this section are quite useful. They follow easily from the singular value decomposition discussed in Section 2.9.

Proposition 2.9. *If A is an $m \times n$ matrix, then*

$$r \leq \min(m, n)$$

We also have the very useful counting rule:

Proposition 2.10. *Let A be an $m \times n$ matrix. It follows*

$$r + \dim \mathcal{N}(A) = n \tag{2.2}$$

Proposition 2.11. *Let A be a real $m \times n$ matrix. Then $rank(A) = rank(A^T) = rank(AA^T) = rank(A^TA)$.*

2.5 PROJECTION MATRICES

The direct sum provides a framework within which a vector space may be systematically split into subspaces that provide a unique expression for the

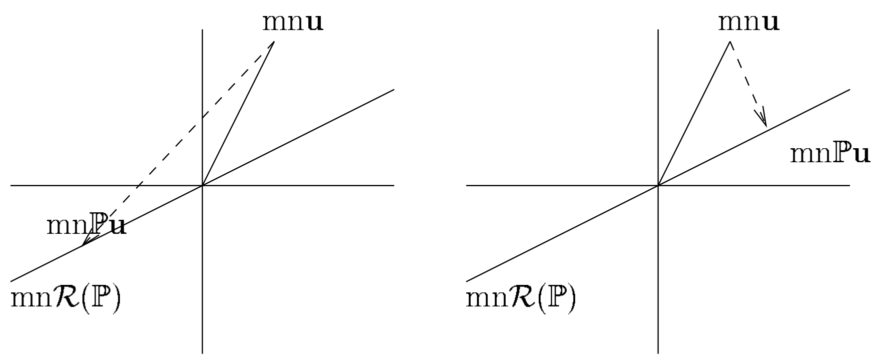

Fig. 2.1 Left: A nonorthogonal, or *oblique* projection. Right: an orthogonal projection.

decomposition of any vector in the space. In this Section we describe a procedure for constructing a mapping which takes a vector and executes this decomposition. Specifically, we refer to a matrix $\mathbb{P}$ as a *projection matrix* if

$$\mathbb{P}^2 = \mathbb{P}.$$

Such matrices are also said to be *idempotent*. See Figure 2.1 which depicts the possible actions of a projection matrix.

Example 2.11. It is easy to verify that the matrix

$$\mathbb{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ 1 & \frac{1}{2} \end{pmatrix}$$

is a projection matrix. Note that it has rank 1 and that

$$\mathcal{R}(\mathbb{P}) = \{\alpha \begin{pmatrix} 1 \\ 2 \end{pmatrix} : \alpha \in \mathbb{R}\}$$

2.5.1 Invariant Subspaces

Definition 2.10. *Let V be a vector space and L a linear operator on V. If W is a subspace of V, we say W is invariant under L if for each $\mathbf{w} \in W$ we have $L\mathbf{w} \in W$. In other words $L(W) \subseteq W$.*

If W_1 and W_2 are subspaces invariant under A (where A is the matrix that corresponds to the linear operator L) with $V = W_1 \oplus W_2$ then we say A is *reduced* or *decomposed* by W_1 and W_2.

We now show that a projection matrix naturally produces an invariant subspace.

Proposition 2.12.

$$\mathbf{v} \in \mathcal{R}(\mathbb{P}) \quad \textit{if and only if} \quad \mathbb{P}\mathbf{v} = \mathbf{v}$$

Proof. First assume $\mathbf{v} \in \mathcal{R}(\mathbb{P})$, i.e., $\mathbf{v} = \mathbb{P}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^n$. So $\mathbb{P}\mathbf{v} = \mathbb{P}^2\mathbf{x}$ but $\mathbb{P}^2\mathbf{x} = \mathbb{P}\mathbf{x} = \mathbf{v}$ from which we conclude that $\mathbb{P}\mathbf{v} = \mathbf{v}$. To prove the converse, assume that $\mathbb{P}\mathbf{v} = \mathbf{v}$. It follows directly that $\mathbf{v} \in \mathcal{R}(\mathbb{P})$. □

2.5.1.1 The Nullspace of $\mathbb{P}$. What is $\mathcal{N}(\mathbb{P})$? A member of this set is readily seen to be the difference between the vector $\mathbf{v}$ being projected and its projection $\mathbb{P}\mathbf{v}$. If $\mathbf{v} = \mathbf{r} + \mathbb{P}\mathbf{v}$ then we have

$$\mathbf{r} = \mathbf{v} - \mathbb{P}\mathbf{v} \tag{2.3}$$

Projecting this vector $\mathbf{r}$ gives

$$\mathbb{P}\mathbf{r} = \mathbb{P}\mathbf{v} - \mathbb{P}^2\mathbf{v} = \mathbf{0}$$

Thus, if $\mathbf{r} \in \mathcal{R}(I - \mathbb{P})$, then $\mathbf{r} \in \mathcal{N}(\mathbb{P})$. Given this is true for an arbitrary $\mathbf{r}$ it follows that $\mathcal{R}(I - \mathbb{P}) \subset \mathcal{N}(\mathbb{P})$. We may also conclude that the null space is invariant under $\mathbb{P}$.

Factoring the projection matrix in equation (2.3) produces

$$\mathbf{r} = (I - \mathbb{P})\mathbf{v} \tag{2.4}$$

so we see the natural decomposition

$$\mathbf{v} = \mathbb{P}\mathbf{v} + (I - \mathbb{P})\mathbf{v} \tag{2.5}$$

The mapping $I - \mathbb{P} : V \to \mathcal{N}(\mathbb{P})$ which takes an element of V to the nullspace of $\mathbb{P}$ is also a projection matrix, known as the *complementary projection matrix*, since

$$\begin{aligned}(I - \mathbb{P})^2 &= I - 2\mathbb{P} + \mathbb{P}^2 \\ &= I - 2\mathbb{P} + \mathbb{P} \\ &= I - \mathbb{P}\end{aligned}$$

We may also employ the notation $\mathbb{Q} = I - \mathbb{P}$ to represent the projection matrix onto the null space.

It is also true that if $\mathbf{r} \in \mathcal{N}(\mathbb{P})$, then $\mathbf{r} \in \mathcal{R}(\mathbb{Q})$. Namely, if $\mathbb{P}\mathbf{r} = \mathbf{0}$, then $(I - \mathbb{P})\mathbf{r} = \mathbf{r}$ so $\mathbf{r} \in \mathcal{R}(\mathbb{Q})$. Again, since $\mathbf{r}$ is arbitrary, it follows that $\mathcal{N}(\mathbb{P}) \subset \mathcal{R}(\mathbb{Q})$. These results lead is to the following proposition:

Proposition 2.13.

$$\mathcal{R}(\mathbb{Q}) = \mathcal{N}(\mathbb{P}) \tag{2.6}$$

2.5.1.2 Independence

Proposition 2.14.

$$\mathcal{R}(\mathbb{P}) \cap \mathcal{N}(\mathbb{P}) = \{\mathbf{0}\} \tag{2.7}$$

Proof. From equation (2.6) we know $\mathcal{N}(\mathbb{P}) = \mathcal{R}(I - \mathbb{P})$. Let $\mathbf{v} \in \mathcal{R}(I - \mathbb{P})$, i.e., $\mathbf{v} = (I - \mathbb{P})\mathbf{x}$ for some $\mathbf{x}$. So $\mathbb{P}\mathbf{v} = \mathbf{0}$. But, by proposition 2.12, $\mathbf{v} \in \mathcal{R}(\mathbb{P})$ if and only if $\mathbb{P}\mathbf{v} = \mathbf{v}$, hence we conclude $\mathbf{v} = \mathbf{0}$ is the only element common to both $\mathcal{R}(\mathbb{P})$ and $\mathcal{N}(\mathbb{P})$. □

From these results it is now clear that a projection matrix separates a space into the sum of two independent subspaces. We recall that this is exactly the direct sum decomposition so we may write

$$V = \mathcal{R}(\mathbb{P}) \oplus \mathcal{N}(\mathbb{P})$$

It is also interesting to note that for every splitting

$$V = W_1 \oplus W_2$$

there exists a projection operator $\mathbb{P}$ such that

$$\mathcal{R}(\mathbb{P}) = W_1$$

and

$$\mathcal{N}(\mathbb{P}) = W_2$$

For details see [69].

2.6 ORTHOGONAL PROJECTION MATRICES

We have seen that projection matrices permit the decomposition of a space into subspaces. The most useful application of this idea is when the resulting subspaces are orthogonal, i.e., when the projection matrix and its complement produce orthogonal vectors. We begin with a basic definition.

Definition 2.11. *Let $\mathbf{x} = \mathbf{w}_1 + \mathbf{w}_2$ and $\mathbf{w}_1 \in W_1, \mathbf{w}_2 \in W_2$ with $W_1 \perp W_2$. The vector $\mathbf{w}_1$ is called the orthogonal projection of $\mathbf{x}$ onto W_1 and $\mathbf{w}_2$ is called the orthogonal projection of $\mathbf{x}$ onto W_2.*

Associated with an orthogonal projection is the operator, which we now refer to as an *orthogonal projection matrix*, which performs the projection described in the definition above. (Note that the orthogonal projection matrix should not be confused with an orthogonal matrix.)

Definition 2.12. *If the subspaces $\mathcal{R}(\mathbb{P})$ and $\mathcal{N}(\mathbb{P})$ are orthogonal, then the projection matrix $\mathbb{P}$ is said to be an orthogonal projection matrix.*

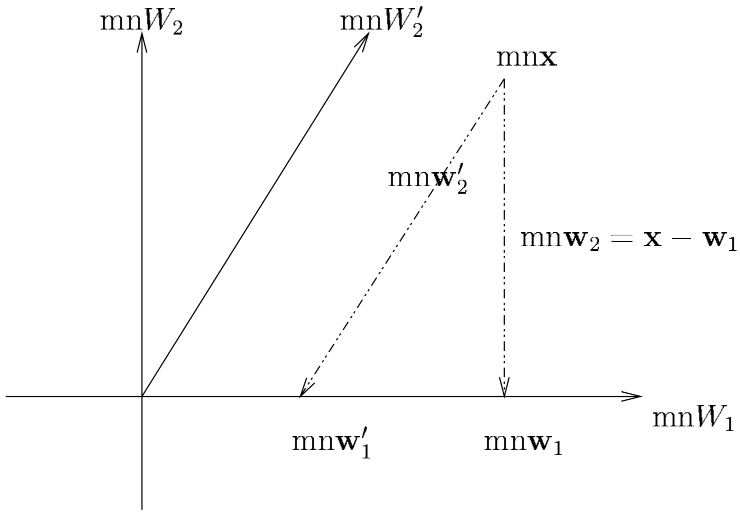

Fig. 2.2 The best approximation to a point $\mathbf{x}$ is the orthogonal projection $\mathbf{w}_1$. Every other projection w_1' has a larger residual w_2'.

If $\mathbb{P}$ is an orthogonal projection matrix, then we may write the direct sum decomposition of the space as

$$V = \mathcal{R}(\mathbb{P}) \dot{\oplus} \mathcal{N}(\mathbb{P}).$$

Why are orthogonal projection matrices to be preferred over plain projection matrices?

2.6.1 Best Approximation Theorem

Suppose W_1 and W_2 are subspaces of an inner product space V s.t. $V = W_1 + W_2$ and let $\mathbf{x} \in V$ be an arbitrary vector. The notion of *best approximation* to $\mathbf{x}$ by a vector in W_1 is made explicit as follows:

Definition 2.13. *A best approximation to* $\mathbf{x}$ *by vectors in* W_1 *is a vector* $\mathbf{w}_1 \in W_1$ *such that*

$$\|\mathbf{x} - \mathbf{w}_1\| \leq \|\mathbf{x} - \mathbf{w}_1'\|$$

for all $\mathbf{w}_1' \in W_1$.

In other words, for each $\mathbf{x} \in V$, we seek a vector $\mathbf{w}_1 \in W_1$ such that $\|\mathbf{x} - \mathbf{w}_1\|$ is a minimum.

Theorem 2.1. The Projection Theorem. *Of all decompositions of the form*

$$\mathbf{x} = \mathbf{w}_1' + \mathbf{w}_2'$$

with $\mathbf{w}_1' \in W_1$*, the orthogonal projection provides the best approximation to* $\mathbf{x}$*. Equivalently, the orthogonal projection minimizes* $\|\mathbf{w}_2'\|$*.*

Proof. We rewrite

$$\begin{aligned}
\|\mathbf{x}-\mathbf{w}_1'\|^2 &= \|\mathbf{x}-\mathbf{w}_1+\mathbf{w}_1-\mathbf{w}_1'\|^2 \\
&= (\mathbf{x}-\mathbf{w}_1+\mathbf{w}_1-\mathbf{w}_1', \mathbf{x}-\mathbf{w}_1+\mathbf{w}_1-\mathbf{w}_1') \\
&= (\mathbf{x}-\mathbf{w}_1, \mathbf{x}-\mathbf{w}_1+\mathbf{w}_1-\mathbf{w}_1') + (\mathbf{w}_1-\mathbf{w}_1', \mathbf{x}-\mathbf{w}_1+\mathbf{w}_1-\mathbf{w}_1') \\
&= (\mathbf{x}-\mathbf{w}_1, \mathbf{x}-\mathbf{w}_1) + (\mathbf{w}_1-\mathbf{w}_1', \mathbf{w}_1-\mathbf{w}_1') + 2(\mathbf{x}-\mathbf{w}_1, \mathbf{w}_1-\mathbf{w}_1') \\
&= \|\mathbf{x}-\mathbf{w}_1\|^2 + \|\mathbf{w}_1-\mathbf{w}_1'\|^2 + 2(\mathbf{x}-\mathbf{w}_1, \mathbf{w}_1-\mathbf{w}_1')
\end{aligned}$$

Observe that

$$\mathbf{x}-\mathbf{w}_1 = \mathbf{w}_2 \in W_2$$

and that $\mathbf{w}_1 - \mathbf{w}_1' \in W_1$. If $W_1 \perp W_2$, i.e., the projection is orthogonal, then it follows that

$$(\mathbf{x}-\mathbf{w}_1, \mathbf{w}_1-\mathbf{w}_1') = 0$$

From this we have

$$\|\mathbf{x}-\mathbf{w}_1'\|^2 \geq \|\mathbf{x}-\mathbf{w}_1\|^2$$

in other words, $\mathbf{w}_1' = \mathbf{w}_1$ is a *best approximation* to $\mathbf{x}$. Note that $\|\mathbf{w}_2'\| = \|\mathbf{x}-\mathbf{w}_1'\|$ is a minimum for $\mathbf{w}_1'$ and since $\mathbf{w}_2 = \mathbf{x}-\mathbf{w}_1$ it follows that $\mathbf{w}_2' = \mathbf{w}_2$ in the case of the best approximation.

Furthermore, it can be shown that this best approximation is unique, see [25] for details. In addition, these results may be extended to the general setting of metric spaces [55].

□

Note that this theorem says nothing about how to select W_1 itself. In other words, given a fixed W_1 the theorem indicates that the orthogonal projection will minimize the error for each vector in V. However, selecting W_1 for a given data set is an entirely different and interesting issue which will be pursued in the sequel.

2.6.2 Criterion for Orthogonal Projections

Proposition 2.15. *If*

$$\mathbb{P} = \mathbb{P}^T \tag{2.8}$$

then the matrix $\mathbb{P}$ *is an orthogonal projection matrix.*

Proof. Let $\mathbb{P} = \mathbb{P}^T$. $\mathbb{P}\mathbf{x} \in \mathcal{R}(\mathbb{P})$ and $(I-\mathbb{P})\mathbf{x} \in \mathcal{N}(\mathbb{P})$.

$$\begin{aligned}
(\mathbb{P}\mathbf{x})^T(I-\mathbb{P})\mathbf{x} &= \mathbf{x}^T\mathbb{P}^T(I-\mathbb{P})\mathbf{x} \\
&= \mathbf{x}^T(\mathbb{P}-\mathbb{P}^2)\mathbf{x} \\
&= \mathbf{0}
\end{aligned}$$

□

The converse of the above proposition is also true, i.e., if $\mathbb{P}$ is an orthogonal projection matrix, then $\mathbb{P} = \mathbb{P}^T$.

Example 2.12. It is easy to verify that the matrix

$$\mathbb{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

is an orthogonal projection matrix. Note that it has rank 1 and that

$$\mathcal{R}(\mathbb{P}) = \alpha \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Example 2.13. Every matrix of the form $\mathbf{v}\mathbf{v}^T$ is an orthogonal projection matrix if $\|\mathbf{v}\| = 1$.

$$\begin{aligned} (\mathbf{v}\mathbf{v}^T)^2 &= (\mathbf{v}\mathbf{v}^T)(\mathbf{v}\mathbf{v}^T) \\ &= \mathbf{v}(\mathbf{v}^T\mathbf{v})\mathbf{v}^T \\ &= \mathbf{v}\mathbf{v}^T \end{aligned}$$

Note that this projection matrix is rank one and that $\mathcal{R}(\mathbf{v}\mathbf{v}^T) = \text{Span}(\mathbf{v})$.

From this example we observe that any vector $\mathbf{u}$ may be orthogonally projected onto a given vector $\mathbf{v}$ by defining

$$\mathbb{P}_{\mathbf{v}}\mathbf{u} = (\mathbf{v}\mathbf{v}^T)\mathbf{u} = \mathbf{v}(\mathbf{v}^T\mathbf{u})$$

Also, the orthogonal complement, or residual $\mathbf{r}$ is then found to be

$$\begin{aligned} \mathbf{r} = \mathbb{P}_{\mathbf{v}}^{\perp} &= (I - \mathbb{P}_{\mathbf{v}})\mathbf{u} \\ &= \mathbf{u} - (\mathbf{v}^T\mathbf{u})\mathbf{v} \end{aligned}$$

We can leverage our ability to project $\mathbf{u}$ onto a single vector $\mathbf{v}$ into a method for computing the orthogonal projection of $\mathbf{u} \in \mathbb{R}^n$ onto a subspace W. To begin, we assume that we have an o.n. basis for the space $\mathbf{W}$ consisting of the vectors $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}\}$. We may view each of the $\mathbf{w}^{(i)}$ as spanning a one-dimensional subspace W_i. Clearly, each of these spaces is orthogonal, i.e.,

$$W_i \perp W_j, \quad i \neq j$$

Furthermore, the sum of these subspaces spans a k-dimensional subspace

$$W = W_1 + \cdots + W_k$$

From our previous deliberations,

$$W = W_1 \dot{\oplus} \ldots \dot{\oplus} W_k$$

In other words, the o.n. basis induces a direct sum decomposition of the subspace W. A projection onto W may be constructed from projections onto the individual supspaces.

The projection of $\mathbf{u}$ onto the i'th subspace space is given by

$$\mathbb{P}_{\mathbf{w}^{(i)}}\mathbf{u} = \mathbf{w}^{(i)}{\mathbf{w}^{(i)}}^T\mathbf{u}$$

If we write $\mathbb{P}_i \equiv \mathbb{P}_{\mathbf{w}^{(i)}}$, then the projection matrix onto W is given by

$$\mathbb{P} = \sum_{i=1}^{k} \mathbb{P}_i = \sum_{i=1}^{k} \mathbf{w}^{(i)}{\mathbf{w}^{(i)}}^T \tag{2.9}$$

Given the matrix $M = [\mathbf{w}^{(1)}|\ldots|\mathbf{w}^{(k)}]$, it follows

$$\mathbb{P} = MM^T. \tag{2.10}$$

2.6.3 Orthogonalization

In the course of the above computations we assumed that the subspace on which we were to project was equipped with an orthonormal basis. We now review the *Gram-Schmidt* procedure for computing an o.n. basis starting from a set of vectors $\{\mathbf{v}^{(i)}\}_{i=1}^{m}$. Take as the first element

$$\mathbf{u}^{(1)} = \frac{\mathbf{v}^{(1)}}{\|\mathbf{v}^{(1)}\|} \tag{2.11}$$

The second element of this set is constructed using the same 2-to-1-dimensional projection technique discussed previously. The projection of $\mathbf{v}^{(2)}$ onto $\mathbf{u}^{(1)}$ is given by

$$\mathbb{P}_{\mathbf{u}^{(1)}}\mathbf{v}^{(2)} = (\mathbf{u}^{(1)}{\mathbf{u}^{(1)}}^T)\mathbf{v}^{(2)}$$

so the vector pointing orthogonally to $\mathbf{u}^{(1)}$ is the residual

$$\mathbf{r} = (I - \mathbb{P}_{\mathbf{u}^{(1)}})\mathbf{v}^{(2)}$$

Simplifying and normalizing this vector gives

$$\mathbf{u}^{(2)} = \frac{\mathbf{v}^{(2)} - ({\mathbf{u}^{(1)}}^T\mathbf{v}^{(2)})\mathbf{u}^{(1)}}{\|\mathbf{v}^{(2)} - ({\mathbf{u}^{(1)}}^T\mathbf{v}^{(2)})\mathbf{u}^{(1)}\|} \tag{2.12}$$

Proceeding in the same fashion with the j'th direction we have

$$\mathbf{u}^{(j)} = \frac{\mathbf{v}^{(j)} - \sum_{i=1}^{j-1}(\mathbf{v}^{(j)}, \mathbf{u}^{(i)})\mathbf{u}^{(i)}}{\|\mathbf{v}^{(j)} - \sum_{i=1}^{j-1}(\mathbf{v}^{(j)}, \mathbf{u}^{(i)})\mathbf{u}^{(i)}\|}$$

Note that if the added direction $\mathbf{v}^{(j)}$ is dependent on the previous vectors then $\mathbf{u}^{(j)} = \mathbf{0}$.

Example 2.14. Consider the matrix

$$A = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

Find the orthogonal projection matrix which takes an element of $\mathbb{R}^4$ onto $\mathcal{R}(A)$. Define $\mathbf{a}^{(1)} = (1\,0\,1\,0)^T$ and $\mathbf{a}^{(2)} = (1\,0\,0\,1)^T$. Given the 3rd column is a multiple of the first $\mathcal{R}(A) = \mathrm{Span}(\mathbf{a}^{(1)}, \mathbf{a}^{(2)})$. To find the projection matrix $\mathbb{P}$ which maps an element of $\mathbb{R}^4$ onto $\mathcal{R}(A)$ we first determine an orthonormal basis for $\mathcal{R}(A)$. Clearly the columns $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ are linearly independent but they are not orthogonal. Using the Gram-Schmidt procedure we obtain

$$\mathbf{u}^{(1)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix}^T$$

and

$$\mathbf{u}^{(2)} = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 & 0 & -1 & 2 \end{pmatrix}^T$$

The projection matrix onto $\mathbf{u}^{(1)}$ is given by

$$\mathbb{P}_1 = \mathbf{u}^{(1)}\mathbf{u}^{(1)T} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and the projection matrix onto $\mathbf{u}^{(2)}$ is given by

$$\mathbb{P}_2 = \mathbf{u}^{(2)}\mathbf{u}^{(2)T} = \frac{1}{6} \begin{pmatrix} 1 & 0 & -1 & 2 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & -2 \\ 2 & 0 & -2 & 4 \end{pmatrix}$$

From this we have the projection matrix

$$\mathbb{P} = \mathbb{P}_1 + \mathbb{P}_2 = \frac{1}{3} \begin{pmatrix} 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 2 & -1 \\ 1 & 0 & -1 & 2 \end{pmatrix}$$

2.7 APPLICATION: THE NOVELTY FILTER

We have seen how projection matrix may be constructed from an arbitrary collection of vectors which span a vector subspace. Now we consider a direct application of these ideas to a pattern processing problem.

Given a set of data set consisting of an ensemble of pattern vectors, e.g., digital images of human faces, we generate associated column vectors by concatenating the rows. In other words, each pattern is available as an n-tuple. Further, let's assume that we are given a large number k of these images but that $k < n$, probably much less. Thus we have an ensemble $\{\mathbf{v}^{(i)}\}_{i=1}^{k}$ where $\mathbf{v}^{(i)} \in \mathbb{R}^n$ for every i.

We would like to determine a projection matrix which takes a new pattern and splits it into two components: the first component is the portion of the data which resides in the subspace spanned by the original patterns, or *training set*; the second component is orthogonal to the training and represents the portion of the data which is *novel*.

With this in mind we define W as the basis in which all the training patterns lie and that $\dim W = m \leq k$ with equality if the original patterns are independent. To determine an orthonormal basis for W the Gram-Schmidt procedure is applied to the training data. This operation will take us from the set of generally non-orthogonal and possibly linearly dependent pattern vectors to an orthonormal basis for W, which we write as the set $\{\mathbf{u}^{(i)}\}_{i=1}^{m}$. In summary,

$$W = \text{Span}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}) = \text{Span}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(m)}).$$

Next, the orthogonal projection matrix $\mathbb{P}$ is computed via equation (2.9), as well as the complementary orthogonal projection matrix $I - \mathbb{P}$. The projection of a pattern produces a point in $\mathbb{R}^m$

$$\mathbb{P} : \mathbb{R}^n \to W$$

$$\mathbf{x} \rightsquigarrow \mathbb{P}\mathbf{x} = \mathbf{w} \in \mathbb{R}^m$$

and the residual sits in $\mathbb{R}^{n-m}$

$$\mathbf{I} - \mathbb{P} : \mathbb{R}^n \to W^\perp$$

$$\mathbf{x} \rightsquigarrow (\mathbf{I} - \mathbb{P})\mathbf{x} = \mathbf{w}^\perp \in \mathbb{R}^{n-m}.$$

As before this is an orthogonal decomposition of

$$\mathbf{x} = \mathbf{w} + \mathbf{w}^\perp.$$

Again, following Kohonen [41], we refer to this orthogonal component as the *novelty* of the pattern, and the general procedure of separating the novelty of a pattern from the non-novel component as the *novelty filter*. In the face data example, novelty might correspond to a new face, or possibly a new pose of a training face.

In practice, problems may arise which make the interpretation of the novelty of a pattern more challenging. Firstly, if the original set of patterns does not include samples of all possible normal patterns, or at least enough to span this set, then the subspace m will be too small and components of a pattern

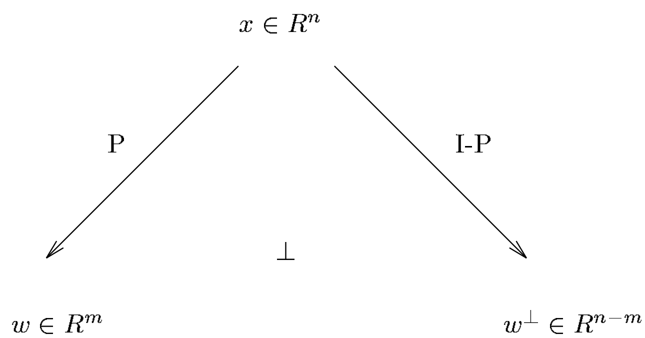

Fig. 2.3 The Novelty Filter Decomposition.

may appear novel only because the set of stored patterns is too small. In addition, the effect of noise on such a subspace representation can be significant. The reader is referred to Kohonen for more details and applications of the novelty filter [41].

2.8 EIGENVALUES AND EIGENVECTORS

Let A be an $n \times n$ matrix. A non-zero vector $\mathbf{v} \in \mathbb{R}^n$ is called an eigenvector and λ its associated eigenvalue if

$$A\mathbf{v} = \lambda \mathbf{v} \tag{2.13}$$

It follows that the only non-trivial solutions to equation (2.13) are obtained when

$$p(\lambda) = \det(A - \lambda I) = 0 \tag{2.14}$$

The determinant may be expanded to produce the *characteristic polynomial*

$$p(\lambda) = (\lambda - \lambda^{(1)})(\lambda - \lambda^{(2)}) \cdots (\lambda - \lambda^{(n)}) \tag{2.15}$$

It is certainly possible that some of the $\lambda^{(i)}$ are the same. The number of times a particular eigenvalue is repeated in equation (2.15) is referred to as its algebraic multiplicity.

Example 2.15. The matrix

$$A = \begin{pmatrix} 1 & -6 & 1 \\ 0 & -3 & -15 \\ 0 & 0 & -3 \end{pmatrix}$$

has the characteristic polynomial

$$p(\lambda) = (\lambda - 1)(\lambda + 3)^2$$

from which we conclude that $\lambda = 1$ is and eigenvalue with (algebraic) multiplicity 1 and $\lambda = -3$ is and eigenvalue with (algebraic) multiplicity 2.

Proposition 2.16. *Every eigenvalue has associated with it at least one eigenvector.*

Proof. Given $\det(A - \lambda I) = 0$

$$\text{rank}(A - \lambda I) < n$$

from which it follows, using equation (2.2), that

$$\dim \mathcal{N}(A - \lambda I) \geq 1.$$

The elements of this nontrivial space are eigenvectors. □

Proposition 2.17. *The eigenvectors associated with the eigenvalue λ, and the zero vector, form a vector subspace which we refer to as the eigenspace E_λ.*

Proof. If $\mathbf{u}, \mathbf{v} \in E_\lambda$ then $A\mathbf{u} = \lambda\mathbf{u}$ and $A\mathbf{v} = \lambda\mathbf{v}$. Let $\mathbf{w} = \alpha\mathbf{u} + \beta\mathbf{v}$.

$$\begin{aligned} A(\alpha\mathbf{u} + \beta\mathbf{v}) &= \alpha A\mathbf{u} + \beta A\mathbf{v} \\ &= \lambda(\alpha\mathbf{u} + \beta\mathbf{v}) \end{aligned}$$

from which we may conclude that $\mathbf{w} \in E_\lambda$. Recall every subspace must contain the zero vector, yet zero is not an eigenvector. □

The eigenspace E_λ is an invariant subspace, i.e., $e \in E_\lambda$ implies $Ae \in E_\lambda$.

Definition 2.14. *The dimension of the eigenspace, i.e., dim(E_λ), is the number of independent eigenvectors associated with λ. This number is also referred to as the geometric multiplicity of λ.*

Proposition 2.18. *The algebraic multiplicity of λ is greater or equal to the geometric multiplicity.*

See [69] for a proof.

An eigenvalue whose geometric multiplicity is less than its algebraic multiplicity is said to be *defective.* An $n \times n$ matrix which has no defective eigenvalues must have n independent eigenvectors.

Theorem 2.2. *Let A be an $n \times n$ matrix with n independent eigenvectors $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \cdots, \mathbf{v}^{(n)}\}$. Define the matrix $V = [\mathbf{v}^{(1)}| \cdots |\mathbf{v}^{(n)}]$. Then*

$$V^{-1}AV = \Lambda$$

where $\Lambda = diag(\lambda^{(1)}, \ldots, \lambda^{(n)})$.

Proof. We will show that $AV = V\Lambda$.

$$\begin{aligned} AV &= A[\mathbf{v}^{(1)}|\cdots|\mathbf{v}^{(n)}] \\ &= [A\mathbf{v}^{(1)}|\cdots|A\mathbf{v}^{(n)}] \\ &= [\lambda^{(1)}\mathbf{v}^{(1)}|\cdots|\lambda^{(n)}\mathbf{v}^{(n)}] \\ &= V\Lambda \end{aligned}$$

Note that the independence of the vectors of V is requires so that V^{-1} exists. □

As a consequence of this theorem, a matrix which has n-independent eigenvectors is said to be *diagonalizable.*

Proposition 2.19. *Eigenvectors associated with distinct eigenvalues are linearly independent.*

Proposition 2.20. *An* $n \times n$ *matrix with* n *distinct eigenvalues*

$$\lambda^{(1)} > \lambda^{(2)} > \cdots > \lambda^{(n)}$$

is diagonalizable.

This follows directly from the fact that the eigenvectors must be independent.

If the eigenbasis is o.n., then the *square* matrix V is said to be orthogonal[1], i.e.,

$$V^T V = I$$

Given V is invertible, it follows that $V^T V = I$ also implies that $V^{-1} = V^T$ and $VV^T = I$. Orthogonal transformations are especially important given they preserve distances in the 2-norm. If V is an orthogonal matrix we have

$$\begin{aligned} \|V\mathbf{x}\|_2^2 &= (V\mathbf{x})^T V\mathbf{x} \\ &= \mathbf{x}^T V^T V\mathbf{x} \\ &= \mathbf{x}^T\mathbf{x} \\ &= \|\mathbf{x}\|_2^2 \end{aligned}$$

Given that the distances are preserved we view the action of an orthogonal matrix as a rotation of the space. Note that this is a special property of the 2-norm.

[1]A matrix is orthogonal simply if its columns are o.n., there is no requirement that they be eigenvectors.

Definition 2.15. *We refer to A as being orthogonally diagonalizable if*

$$V^T A V = \Lambda$$

Theorem 2.3. *Spectral Theorem: The matrix A is symmetric iff there is a real orthogonal matrix V s.t.*

$$V^T A V = \Lambda$$

For a proof of this theorem the reader is referred to [26].

The equation $V^T A V = \Lambda$ may be rewritten as

$$A = V \Lambda V^T = \sum_i \lambda^{(i)} \mathbf{v}^{(i)} \mathbf{v}^{(i)T}$$

This representation expresses a square matrix in terms of a sum of rank one matrices.

The proof of the theorem relies on two basic facts which are proved in elementary linear algebra courses. We bundle them in a single proposition.

Proposition 2.21. *Let A be an $n \times n$ symmetric matrix.*

- *The eigenvalues of A are real.*
- *If $\lambda^{(i)} \neq \lambda^{(j)}$ then the eigenvectors $\mathbf{v}^{(i)}$ and $\mathbf{v}^{(j)}$ are orthogonal.*
- *A is not defective, i.e., it has n independent eigenvectors.*

Thus, given an $n \times n$ symmetric matrix A, an o.n. basis for $\mathbb{R}^n$ may be constructed from its eigenvectors. Eigenvalues of algebraic multiplicity one have orthogonal eigenvectors; eigenvectors which correspond to an eigenvalue with multiplicity greater than one may be orthogonalized by applying the Gram-Schmidt procedure. Note that for symmetric matrices the algebraic and geometric multiplicity of any eigenvalue must always be the same.

Change of Basis Revisited I

Let A be an $n \times n$ real matrix. The linear system

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

may be rewritten in an especially simple form if A is diagonalizable. Assume the matrix V diagonalizes A, i.e., $\Lambda = V^{-1} A V$. Introduce the change of coordinates

$$\mathbf{x} = V\mathbf{x}' \quad \mathbf{b} = V\mathbf{b}'$$

Then $A\mathbf{x} = \mathbf{b}$ becomes $AV\mathbf{x}' = V\mathbf{b}'$ or

$$\Lambda \mathbf{x}' = \mathbf{b}'.$$

2.9 THE SINGULAR VALUE DECOMPOSITION

The singular value decomposition (SVD) extends the spectral theorem for rectangular matrices. We shall see in Chapter 3.1 that it also provides the necessary mathematics for understanding an important class of optimal dimensionality reducing mappings. We shall begin with a statement of the decomposition theorem and in the course of proving it, we will establish several important facts concerning the SVD. The early history of the SVD is recounted in [66]. A detailed account of the theory is available in [27].

Theorem 2.4. *Singular Value Decomposition (SVD). Let A be a real $m \times n$ matrix and $l = \min\{m, n\}$. There exist orthogonal matrices U and V such that*

$$A = U\Sigma V^T \tag{2.16}$$

where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ and $\Sigma = diag(\sigma^{(1)}, \dots, \sigma^{(l)}) \in \mathbb{R}^{m\times n}$. Futhermore, the entries of Σ are ordered according to

$$\sigma^{(1)} \geq \sigma^{(2)} \geq \cdots \geq \sigma^{(l)} \geq 0.$$

The case for A being a complex matrix is analogous and is treated, e.g., in [69]. Without loss of generality we may assume that $m \geq n$ in what follows.

To establish this decomposition we first rewrite it using

$$AV = U\Sigma$$

The i'th column of this relationship is

$$A\mathbf{v}^{(i)} = \sigma^{(i)}\mathbf{u}^{(i)} \tag{2.17}$$

where $i = 1, \dots, l$. Alternatively,

$$A^T U = V\Sigma^T$$

The i'th column of this relationship is

$$A^T\mathbf{u}^{(i)} = \sigma^{(i)}\mathbf{v}^{(i)} \tag{2.18}$$

where again $i = 1, \dots, l$. The solutions of these equations occur in triples $\{\sigma^{(i)}, \mathbf{u}^{(i)}, \mathbf{v}^{(i)}\}$ consisting of the singular values $\sigma^{(i)}$, left-singular vectors $\mathbf{u}^{(i)}$ and the right-singular vectors $\mathbf{v}^{(i)}$.

Proposition 2.22. *The left-singular vectors of A are eigenvectors of AA^T and the associated eigenvalues correspond to the singular values squared.*

$$A^T\mathbf{u} = \sigma\mathbf{v}$$

$$AA^T\mathbf{u} = \sigma A\mathbf{v}$$
$$AA^T\mathbf{u} = \sigma^2\mathbf{u}$$

and hence, $\sqrt{\lambda} = \sigma$.

Note that the size of this eigenvector problem is $m \times m$.

Proposition 2.23. *The left-singular vectors of A, augmented by the eigenvectors in the null-space AA^T, form an o.n. basis for $\mathbb{R}^m$.*

In other words,

$$U = [\mathbf{u}^{(1)}| \cdots |\mathbf{u}^{(n)}|\mathbf{u}^{(n+1)}| \cdots |\mathbf{u}^{(m)}]$$

where $\hat{U} = [\mathbf{u}^{(1)}| \cdots |\mathbf{u}^{(n)}]$ is the submatrix whose columns are simultaneously left-singular vectors of A *and* eigenvectors of AA^T, while the remaining $\mathbf{u}^{(n+1)}| \cdots |\mathbf{u}^{(m)}]$ are eigenvectors in the nullspace of AA^T.

An analogous proposition is true for the right-singlar vectors.

Proposition 2.24. *The right-singular vectors of A are the eigenvectors of A^TA and the associated eigenvalues correspond to the singular values squared.*

$$A\mathbf{v} = \sigma\mathbf{u}$$
$$A^TA\mathbf{v} = \sigma A^T\mathbf{u}$$
$$A^TA\mathbf{v} = \sigma^2\mathbf{v}$$

Note that this is a $n \times n$ eigenvector problem. However, since we are assuming $n \leq m$, all of the eigenvectors in this instance are also singular vectors.

Proposition 2.25. *The right-singular vectors of A form an o.n. basis for $\mathbb{R}^m$.*

In other words,

$$V = [\mathbf{v}^{(1)}| \cdots |\mathbf{v}^{(n)}]$$

Proposition 2.26. *If rank$(A) = r$, then there are r non-zero singular values, i.e.,*

$$\sigma^{(1)} \geq \cdots \geq \sigma^{(r)} > \sigma^{(r+1)} = 0$$

This result may be established in a variety of ways.

- The rank of a diagonal matrix is the number of non-zero diagonal elements. Furthermore, orthogonal transformations do not change the number of vectors that make up a basis. In view of $A = U\Sigma V^T$, it follows that if rank$(\Sigma) = r$ then rank$(A) = r$.
- The column space has r dimensions and the left-singular vectors corresponding to non-zero singular vectors form a basis for this space.

- A counting argument using equation (2.23) establishes the size of the nullspace, and hence rank of A. See exercise 2.25.

We are now in a position to provide a constructive proof of the SVD based on the existence of the left and right singular vectors. Again, we assume $m \geq n$, so there are n singular vector triplets $\{\sigma^{(i)}, \mathbf{u}^{(i)}, \mathbf{v}^{(i)}\}$. First we will show that

$$\underset{m\times n}{A}\underset{n\times n}{V} = \underset{m\times n}{\hat{U}}\underset{n\times n}{\hat{\Sigma}}$$

$$\begin{aligned} AV &= A[\mathbf{v}^{(1)}|\cdots|\mathbf{v}^{(n)}] \\ &= [A\mathbf{v}^{(1)}|\cdots|A\mathbf{v}^{(n)}] \\ &= [\sigma^{(1)}\mathbf{u}^{(1)}|\cdots|\sigma^{(n)}\mathbf{u}^{(n)}] \\ &= \hat{U}\hat{\Sigma} \end{aligned}$$

It is sometimes convenient to rewrite this decomposition. For example, if $m \geq n$ then we may write

$$A = \hat{U}\hat{\Sigma}V^T \tag{2.19}$$

where $\hat{U} \in \mathbb{R}^{m\times n}$ (i.e., U with the last $m-n$ columns deleted), and $\hat{\Sigma} = \mathbb{R}^{n\times n}$. This version of the SVD is referred to as the *thin* SVD [20], or the *reduced* SVD [69].

The full SVD follows by including the eigenvectors $\{\mathbf{u}^{(n+1)}, \ldots, \mathbf{u}^{(m)}\}$. In this case we write

$$AV = [\mathbf{u}^{(1)}|\cdots|\mathbf{u}^{(n)}|\mathbf{u}^{(n+1)}|\cdots|\mathbf{u}^{(m)}]\begin{pmatrix} \sigma^{(1)} & & 0 \\ & \ddots & \\ 0 & & \sigma^{(n)} \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \tag{2.20}$$

The singular value decomposition provides bases for the fundamental subspaces. We allow the matrix A to be *rank deficient*, i.e., rank $r \leq l = \min\{m, n\}$.

Proposition 2.27. *Let A be an $m \times n$ matrix of rank r. Then:*

1. *The r left-singular vectors $\{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(r)}\}$ corresponding to non-zero singular values form a basis for $\mathcal{R}(A)$.*

2. *The $m-r$ vectors $\{\mathbf{u}^{(r+1)}, \ldots, \mathbf{u}^{(m)}\}$ (the first $n-r$ are left-singular vectors) corresponding to eigenvectors in $\mathcal{N}(AA^T)$ form a basis for $\mathcal{N}(A^T)$.*

3. *The r right-singular vectors of A $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(r)}\}$ form a basis for $\mathcal{R}(A^T)$.*

4. The $n-r$ *right-singular vectors of* A $\{\mathbf{v}^{(r+1)}, \ldots, \mathbf{v}^{(n)}\}$ *form a basis for* $\mathcal{N}(A)$.

These bases may be employed to form direct sum decompositions of $\mathbb{R}^m$ and $\mathbb{R}^n$ as

$$\mathbb{R}^m = \mathcal{R}(A)\dot{\oplus}\mathcal{N}(A^T)$$

and

$$\mathbb{R}^n = \mathcal{R}(A^T)\dot{\oplus}\mathcal{N}(A).$$

Furthermore, we may conclude that

$$r + \dim \mathcal{N}(A^T) = m$$

and

$$r + \dim \mathcal{N}(A) = n.$$

Example 2.16. Compute the SVD of the data matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

First, we compute the right-singular vectors and singular values of A. These are exactly the eigenvectors and the square roots of the eigenvalues of A^TA,

$$A^TA = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

which has the characteristic equation

$$p(\lambda) = (\lambda - 1)(\lambda - 3) = 0$$

Hence $\sigma^{(1)} = \sqrt{\lambda^{(1)}} = \sqrt{3}$ has the right-singular vector (eigenvector)

$$\mathbf{v}^{(1)} = \frac{1}{\sqrt{2}}(1, 1)^T$$

and $\sigma^{(2)} = \sqrt{\lambda^{(2)}} = 1$ has the right singular vector (eigenvector)

$$\mathbf{v}^{(2)} = \frac{1}{\sqrt{2}}(-1, 1)^T$$

So the matrix of right singular vectors is

$$V = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

The 2 left-singular vectors are given by the eigenvectors of

$$AA^T = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

corresponding to the 2 largest eigenvalues. The characteristic equation for AA^T is given by

$$p(\lambda) = \lambda(\lambda - 1)(\lambda - 3) = 0$$

Now $\sigma^{(1)}$ (seen to be the same as above, as expected) has the left-singular vector

$$\mathbf{u}^{(1)} = \frac{1}{\sqrt{6}}(2, 1, 1)^T$$

and $\sigma^{(2)} = \sqrt{\lambda^{(2)}} = 1$ has the left singular vector

$$\mathbf{u}^{(2)} = \frac{1}{\sqrt{2}}(0, 1, -1)^T$$

The eigenvalue $\lambda^{(3)} = 0$ corresponds to the *eigenvector* $\mathbf{u}^{(3)} = \frac{1}{\sqrt{3}}(1, -1, -1)^T$. This last vector completes the basis for $\mathbb{R}^3$.

$$U = \begin{pmatrix} 2/\sqrt{6} & 0 & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & -1/\sqrt{3} \end{pmatrix}$$

So the full SVD decomposition is given by

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

and the reduced SVD decomposition is given by

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

2.9.1 Reduction and Compression of Matrices

The SVD may be used to provide an efficient representation of a rank r matrix. Without loss of information it decomposes, or reduces, the matrix into a sum of r rank one matrices. Specifically, given that the matrix A has rank r it follows from equation (2.16) that

$$A = \sum_{j=1}^{r} \sigma^{(j)} \mathbf{u}^{(j)} \mathbf{v}^{(j)^T} \tag{2.21}$$

To see this, it is useful to define the matrices $\Sigma_1 = \mathrm{diag}(\sigma^{(1)}, 0, \ldots, 0)$, $\Sigma_2 = \mathrm{diag}(0, \sigma^{(2)}, 0 \ldots, 0)$, and so on. It follows

$$\begin{aligned} A &= U\Sigma V^T \\ &= U(\Sigma_1 + \Sigma_2 + \cdots + \Sigma_r)V^T \\ &= U\Sigma_1 V^T + U\Sigma_2 V^T + \cdots + U\Sigma_r V^T \\ &= \sigma^{(1)}\mathbf{u}^{(1)}\mathbf{v}^{(1)}{}^T + \sigma^{(2)}\mathbf{u}^{(2)}\mathbf{v}^{(2)}{}^T + \cdots + \sigma^{(r)}\mathbf{u}^{(r)}\mathbf{v}^{(r)}{}^T \end{aligned}$$

The SVD provides not only an efficient means to represent a matrix without loss, it also provides an optimal method for appoximating a matrix by another matrix of reduced rank. Define a rank $k < r$ approximation to the matrix A as

$$A_k = \sum_{j=1}^{k} \sigma^{(j)}\mathbf{u}^{(j)}\mathbf{v}^{(j)}{}^T \tag{2.22}$$

Proposition 2.28.

$$\|A - A_k\|_2 = \sigma^{(k+1)}$$

Proof.

$$\begin{aligned} A - A_k &= \sum_{j=1}^{r} \sigma^{(j)}\mathbf{u}^{(j)}\mathbf{v}^{(j)}{}^T - \sum_{j=1}^{k} \sigma^{(j)}\mathbf{u}^{(j)}\mathbf{v}^{(j)}{}^T \\ &= \sum_{j=k+1}^{r} \sigma^{(j)}\mathbf{u}^{(j)}\mathbf{v}^{(j)}{}^T \\ &= U\Sigma' V^T \end{aligned}$$

where

$$\Sigma' = \begin{pmatrix} 0 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 0 & & & & & & \\ & & & \sigma^{(k+1)} & & & & & \\ & & & & \ddots & & & & \\ & & & & & \sigma^{(r)} & & & \\ & & & & & & 0 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 0 \end{pmatrix}$$

It follows that

$$\begin{aligned} \|A - A_k\|_2 &= \|U\Sigma' V^T\|_2 \\ &= \|\Sigma'\|_2 \\ &= \sigma^{(k+1)} \end{aligned}$$

since 2-norms are invariant under multiplication by orthogonal matrices.
□

The SVD provides the best reduced rank approximation to a given matrix A. Any matrix, say B, which is not the rank k SVD approximation has greater error.

Theorem 2.5.

$$A_k = \arg \min_{rank(B)=k} \|A - B\|_2$$

For a proof of this important theorem see [20].

Example 2.17. A rank one approximation to the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \approx \begin{pmatrix} 1 & 1 \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

This approximation is calculated by

$$A \approx A_1 = \sigma^{(1)} \mathbf{u}^{(1)} \mathbf{v}^{(1)T}$$

2.9.2 Applications of the SVD

The SVD provides a powerful means to approximate a matrix with minimum error. In addition, as we have already seen, it provides a means to explore the properties of these matrices. We continue with a few more applications.

Proposition 2.29. *Let A be an $m \times n$ matrix of rank r. Then $r = rank(AA^T) = rank(A^T A)$.*

This follows directly from the correspondence of the non-zero singular values with the non-zero eigenvalues: they are exactly the same in number.

The SVD permits an nice geometric interpretation of the application of a matrix to the hypersphere $\|\mathbf{x}\|_2 = 1$.

$$A\mathbf{x} = U \Sigma V^T \mathbf{x}$$

The multiplication term $\mathbf{y} = V^T \mathbf{x}$ just rotates the hypersphere so $\|\mathbf{y}\| = 1$. The second multiplication $\mathbf{z} = \Sigma \mathbf{y}$ maps the hypersphere a hyperellipse with semi-axes $\sigma^{(i)}$ and the final multiplication $U\mathbf{z}$ serves to rotate the hypersphere out of the standard basis.

Proposition 2.30. *The singular values of the $m \times n$ matrx A correspond to the lengths of the semi-axes of the hyper-ellipsoid*

$$\{A\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$$

Change of Basis Revisited II

Consider again the problem

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

where $A \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$. We may introduce new coordinate systems for the domain and range of A using the SVD. Let

$$\mathbf{b} = U\mathbf{b}'$$

and

$$\mathbf{x} = V\mathbf{x}'$$

Then $A\mathbf{x} = \mathbf{b}$ becomes

$$U\mathbf{b}' = AV\mathbf{x}'$$

$$\mathbf{b}' = U^T AV\mathbf{x}'$$

Since $U^T AV = \Sigma$ by the SVD we have

$$\mathbf{b}' = \Sigma\mathbf{x}'$$

This shows that every linear problem is diagonal in the coordinate system provided by the SVD!

2.9.3 Computation of the SVD

We have shown that the singular values may be computed by forming the covariance matrices AA^T or $A^T A$ and computing their eigenvalues. While this approach is suitable for many applications, it is numerically unstable [69]. An alternative to forming the covariance matrices is to calculate the left and right singular vectors directly from the system

$$\begin{pmatrix} \mathbf{0} & A \\ A^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u}^{(i)} \\ \mathbf{v}^{(i)} \end{pmatrix} = \sigma^{(i)} \begin{pmatrix} \mathbf{u}^{(i)} \\ \mathbf{v}^{(i)} \end{pmatrix} \tag{2.23}$$

If the matrix A is perturbed by a small amount then it can be shown that the perturbed singular values $\tilde{\sigma}^{(i)}$ satisfy

$$|\tilde{\sigma}^{(i)} - \sigma^{(i)}| = O(\epsilon \|A\|)$$

when computed using equation (2.23); ϵ is the machine precision. On the other hand, if the singular values are obtained by first computing the eigenvalues of the smaller of the two matrices $A^T A$ or AA^T then

$$|\tilde{\sigma}^{(i)} - \sigma^{(i)}| = O(\epsilon \|A\|^2 / \sigma^{(i)})$$

This squaring of norm, followed by the division of the singular value becomes significant especially for the smaller singular values. See [69] for a complete discussion.

Example 2.18. To demonstrate the mechanics of the system calculation we revisit Example 2.16. Now we have to compute the eigenvectors and eigenvalues of

$$\begin{pmatrix} \mathbf{0} & A \\ A^T & \mathbf{0} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

The eigenvectors are

$$\begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{\sqrt{3}} \\ \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 \end{pmatrix}$$

where the columns are ordered from left-to-right via the singular values $\sigma = -\sqrt{3}, \sqrt{3}, -1, 1, 0$. The eigenvectors now contain the right and left singular vectors as components, i.e., each column is of the form $(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})$. It is interesting to note that all the singular values $\pm\sigma^{(i)}$ are present and that the associated eigenvectors are $(\mathbf{u}^{(i)}, \pm\mathbf{v}^{(i)})$. For more details see Problem 2.24.

Problems

2.1 Show that the transformation

$$T\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}\right) = \begin{pmatrix} x_1 + x_2 \\ x_2 + x_3 \\ x_3 + x_4 \\ x_4 + x_1 \end{pmatrix}$$

is linear. Determine the matrix which represents this transformation.

2.2 Using the relationship $\mathbf{v}^{(i)} = \sum_{j=1}^{n} p_{ij}\mathbf{w}^{(j)}$ show that $\mathbf{y} = \mathbf{P}^T\mathbf{x}$ and deduce that $(\mathbf{P}^T)^{-1} = \mathbf{Q}^T$, i.e., the coordinate transformation is invertible.

2.3 Consider the vector $\mathbf{v}$ whose coordinates w.r.t. the basis $\mathcal{B}_1$ defined in Example 2.6 are $(3, 5)$. What are the coordinates of $\mathbf{v}$ w.r.t. $\mathcal{B}_2$?

2.4 Let the basis $\mathcal{B}_1$ be the standard basis, i.e., $\mathbf{e}^{(1)} = (1\,0)^T, \mathbf{e}^{(1)} = (0\,1)^T$ and the basis $\mathcal{B}_2$ be given by the two vectors $\mathbf{v}^{(1)} = (1\,1)^T, \mathbf{v}^{(2)} = (-1\,1)^T$. Given $\mathbf{u}_{\mathcal{B}_1} = (1\,1)^T$ find $\mathbf{u}_{\mathcal{B}_2}$.

2.5 Let $\mathcal{B}_1$ be the standard basis and $\{\mathbf{w}^{(i)}\}$ be the vectors which define $\mathcal{B}_2$. Given $\mathbf{u}_{\mathcal{B}_2} = P^T\mathbf{u}_{\mathcal{B}_1}$ show $P^T = W^{-1}$ where $W = [\mathbf{w}^{(1)}|\ldots|\mathbf{w}^{(n)}]$.

2.6 Let the linear mapping L correspond to multiplication by the matrix A

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix}$$

which is given w.r.t. the basis $\mathcal{B}_1$ made up of the vectors $\mathbf{v}^{(1)} = (1\,1)^T$ and $\mathbf{v}^{(2)} = (1\,-1)^T$. Find the matrix A' which corresponds to the same mapping L but now w.r.t. the basis $\mathcal{B}_2$ made up of the vectors $\mathbf{w}^{(1)} = (1\,0)^T$ and $\mathbf{w}^{(2)} = (1\,2)^T$.

2.7 Define the union of two subspaces. Show that it is generally *not* a subspace.

2.8 Let W_1 and W_2 be vector subspaces and $W = W_1 + W_2$. Show, by giving an example, that the decomposition of a vector $\mathbf{x} \in W$ is not unique, i.e.,

$$\mathbf{x} = \mathbf{w}_1 + \mathbf{w}_2 = \mathbf{w}_1' + \mathbf{w}_2'$$

where $\mathbf{w}_1 \neq \mathbf{w}_1'$, $\mathbf{w}_2 \neq \mathbf{w}_2'$ and $\mathbf{w}_1, \mathbf{w}_1' \in W_1$, $\mathbf{w}_2, \mathbf{w}_2' \in W_2$.

2.9 Let W be a subspace of the vector space V and let W_1 and W_2 be subspaces of W s.t. $W = W_1 + W_2$ and $W_1 \cap W_2 = \{\mathbf{0}\}$. Show directly that any $\mathbf{x} \in W$ can be uniquely decomposed as

$$\mathbf{x} = \mathbf{w}_1 + \mathbf{w}_2$$

where $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$.

2.10 Prove that if $W_1 \cap W_2 = \{\mathbf{0}\}$, then W_1 and W_2 are independent subspaces.

2.11 Consider the statements:

- Two independent subspaces must be orthogonal.
- Two orthogonal subspaces must be independent.

In each case, either prove or provide a counter example.

2.12 Prove that the column space of a matrix is a subspace.

2.13 Are perpendicular planes orthogonal? Are they independent? Either prove, or provide vectors which are either not orthogonal or independent.

2.14 Find the row space $\mathcal{R}(A^T)$ and left null space $\mathcal{N}(A^T)$ for the matrix $A = \mathbf{u}\mathbf{v}^T$.

2.15 Show that

$$\mathcal{R}(\mathbb{P}) = \mathcal{N}(\mathbb{I} - \mathbb{P}) \tag{2.24}$$

where $\mathbb{P}$ is a projection matrix.

2.16 Show that in general,

$$\mathbb{P}_{\mathbf{v}} = \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}$$

and

$$\mathbb{P}_{\mathbf{v}}^{\perp} = I - \frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}$$

2.17 Let $V = \mathbb{R}^3$ and

$$\text{let } \mathbf{u}^{(1)} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} \text{ and } \mathbf{u}^{(2)} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \text{ and let } \mathbf{x} = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}$$

and define $W_1 = \text{Span}(\mathbf{u}^{(1)}, \mathbf{u}^{(2)})$.

Find the orthogonal projection of $\mathbf{x}$ onto W_1. Also find the projection matrix $\mathbb{P}$ associated with this mapping.

2.18 If $\mathbb{P}$ is an orthogonal projection matrix show that $I - \mathbb{P}$ is also an orthogonal projection matrix.

2.19 Are distances preserved by orthogonal transformations in the 1-norm? Recall

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

2.20 Let $G \in \mathbb{R}^{n\times n}$ not be an orthogonal matrix. What is the closest orthogonal matrix to G in the 2-norm?

2.21 The full SVD is written

$$A = U\Sigma V^T$$

and if $m \geq n$ the reduced SVD is written

$$A = \hat{U}\hat{\Sigma}V^T$$

Show that U and V are orthogonal matrices while the reduced matrices $\hat{U}, \hat{\Sigma}$ are not.

2.22 Show that $\|A\mathbf{v}^{(i)}\|_2 = \sigma^{(i)}$ and interpret geometrically.

2.23 Determine the SVD of the data matrix

$$A = \begin{pmatrix} -2 & -1 & 1 \\ 0 & -1 & 0 \\ -1 & 1 & 2 \\ 1 & -1 & 1 \end{pmatrix}$$

and compute the rank one, two and three approximations to A.

2.24 Show that if $\begin{pmatrix} \mathbf{u}^{(i)} & \mathbf{v}^{(i)} \end{pmatrix}^T$ is an eigenvector corresponding to eigenvalue σ of

$$\begin{pmatrix} \mathbf{0} & A \\ A^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u}^{(i)} \\ \mathbf{v}^{(i)} \end{pmatrix} = \sigma^{(i)} \begin{pmatrix} \mathbf{u}^{(i)} \\ \mathbf{v}^{(i)} \end{pmatrix} \tag{2.25}$$

then $\begin{pmatrix} \mathbf{u}^{(i)} & -\mathbf{v}^{(i)} \end{pmatrix}^T$ is an eigenvector corresponding to eigenvalue $-\sigma$

2.25 Let the rank of A be r. Consider the eigenvector system in the previous problem.

- Show that this is a symmetric eigenvector problem.
- How many independent eigenvectors are there of the form

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$$

- How many eigenvectors of the type

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix} \text{ or } \begin{pmatrix} \mathbf{0} \\ \mathbf{v} \end{pmatrix}$$

  Hint: What is the geometric multiplicity of the eigenvalue $\lambda = 0$ for the matrices A and A^T?

- Given that the singular values occur in pairs $\pm\sigma$, determine the number of positive singular values for A.

2.26 The SVD for the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

is provided in Example 2.16. Using the computed decomposition $A = U\Sigma V^T$, describe the action of the matrix A on the unit circle. Specifically identify the image of the unit vectors

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \end{pmatrix}^T \quad \text{and} \quad \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \end{pmatrix}^T$$

Part II

Optimal Orthogonal Pattern Representations