Chapter 2

Vector Spaces

Our first technical topic for this book is linear algebra, which is one of the foundation stones of applied mathematics in general, and econometrics and statistics in particular. Data ordered by observation are naturally stored in vectors. Related vectors are naturally grouped into matrices. Once our data are organized this way, we need to perform basic arithmetic operations or solve equations or quadratic minimization problems. In this chapter we cover the foundations of vector operations used in linear algebra. As we'll see, the conceptual aspects of linear algebra are clearest if we begin by stripping away details such as matrices and look instead at linear operations from a more abstract perspective.

2.1 Vectors and Vector Space

Let's begin with vector space and basic vector operations.

2.1.1 Vectors

For arbitrary $N \in \mathbb{N}$, the symbol $\mathbb{R}^N$ represents the set of all N-vectors, or vectors of length N. A typical element has the form

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \quad \text{where } x_n \in \mathbb{R} \text{ for each } n$$

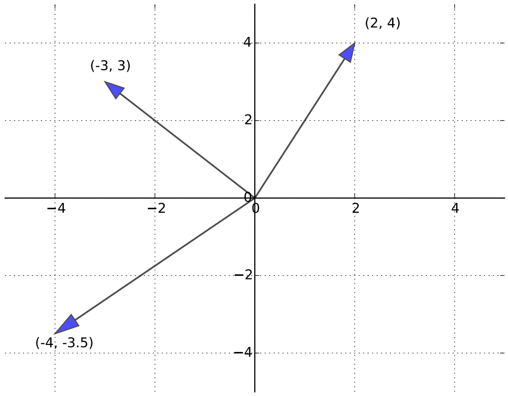


Figure 2.1 Three vectors in $\mathbb{R}^2$

($\mathbb{R} = \mathbb{R}^1$ represents the set of all real numbers, which is the union of the rational and irrational numbers.) While $\mathbf{x}$ has been written vertically, as a column of numbers, we could also write it horizontally as $\mathbf{x} = (x_1, \ldots, x_N)$. For now vectors are just a sequences of numbers, and it makes no difference whether we write them vertically or horizontally. (Only when we work with matrix multiplication does it become necessary to distinguish between column and row vectors.)

The vector of ones will be denoted $\mathbf{1}$ and the vector of zeros will be denoted $\mathbf{0}$:

$$\mathbf{1} := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \mathbf{0} := \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Although vectors are infinitesimal points in $\mathbb{R}^N$, they are often represented visually as arrows from the origin to the point itself. Figure 2.1 gives an illustration for the case $N = 2$.

In vector space theory there are two fundamental algebraic operations: vector addition and scalar multiplication. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, their **vector sum** is

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} := \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_N + y_N \end{pmatrix}$$

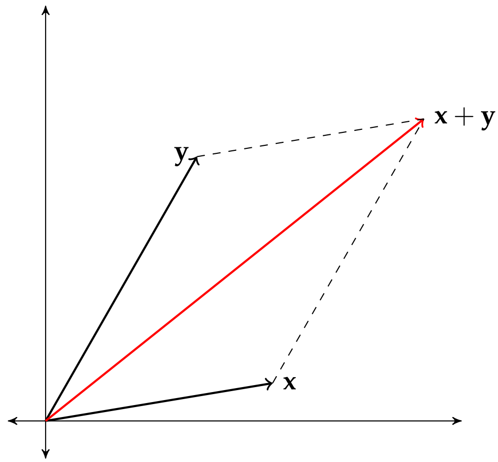


Figure 2.2 Vector addition

If $\alpha \in \mathbb{R}$, then the **scalar product** of α and $\mathbf{x}$ is defined to be

$$\alpha \mathbf{x} = \alpha \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} := \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_N \end{pmatrix}$$

Thus addition and scalar multiplication are defined in terms of ordinary addition and multiplication in $\mathbb{R}$, computed element by element, adding and multiplying respectively.[1] Figures 2.2 and 2.3 show examples of vector addition and scalar multiplication in the case $N = 2$.

Subtraction of two vectors is performed element by element, just like addition. Subtraction is not a primitive operation because the definition can be given in terms of addition and scalar multiplication: $\mathbf{x} - \mathbf{y} := \mathbf{x} + (-1)\mathbf{y}$. An illustration of subtraction is given in figure 2.4. One way to remember this operation is to draw a line from $\mathbf{y}$ to $\mathbf{x}$ and then shift it to the origin.

The **inner product** of two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^N$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$, and defined

1. In some instances, the notion of scalar multiplication includes multiplication of vectors by complex numbers. In what follows we will work almost entirely with real scalars. Hence scalar multiplication means real scalar multiplication unless otherwise stated.

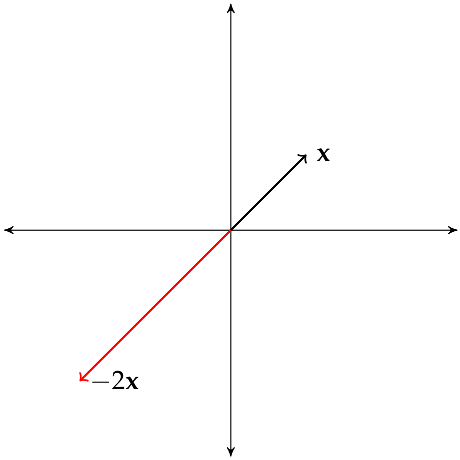


Figure 2.3 Scalar multiplication

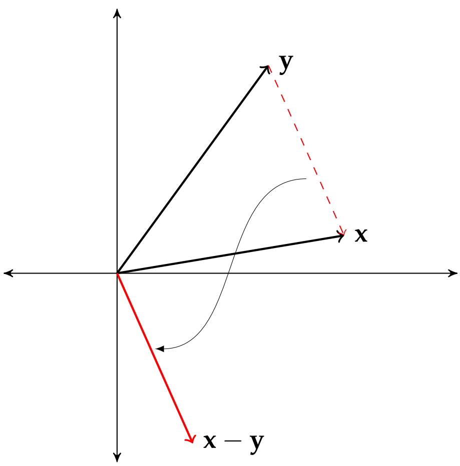


Figure 2.4 Difference between vectors

as the sum of the products of their elements:

$$\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{n=1}^{N} x_n y_n \tag{2.1}$$

Fact 2.1.1 For any $\alpha, \beta \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^N$, the following statements are true:

(i) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$,

(ii) $\langle \alpha\mathbf{x}, \beta\mathbf{y} \rangle = \alpha\beta \langle \mathbf{x}, \mathbf{y} \rangle$, and

(iii) $\langle \mathbf{x}, \alpha\mathbf{y} + \beta\mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle + \beta \langle \mathbf{x}, \mathbf{z} \rangle$.

These properties are easy to check from (2.1). For example, regarding the second equality, pick any $\alpha, \beta \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. By the definitions of scalar multiplication and inner product respectively, we have

$$\langle \alpha\mathbf{x}, \beta\mathbf{y} \rangle = \sum_{n=1}^{N} \alpha x_n \beta y_n = \alpha\beta \sum_{n=1}^{N} x_n y_n = \alpha\beta \langle \mathbf{x}, \mathbf{y} \rangle$$

The (Euclidean) **norm** of a vector $\mathbf{x} \in \mathbb{R}^N$ is defined as

$$\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \tag{2.2}$$

and represents the length of the vector $\mathbf{x}$. (In the arrow representation of vectors in figures 2.2–2.4, the norm of the vector is equal to the length of the arrow.)

Fact 2.1.2 For any $\alpha \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, the following statements are true:

(i) $\|\mathbf{x}\| \geqslant 0$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$,

(ii) $\|\alpha\mathbf{x}\| = |\alpha| \|\mathbf{x}\|$,

(iii) $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$, and

(iv) $|\langle \mathbf{x}, \mathbf{y} \rangle| \leqslant \|\mathbf{x}\| \|\mathbf{y}\|$.

Properties (i) and (ii) you can verify yourself without difficulty. Proofs for (iii) and (iv) are a bit harder. Property (iii) is called the **triangle inequality**, while (iv) is called the **Cauchy–Schwarz inequality**. The proof of the Cauchy–Schwarz inequality is given as a solved exercise after we've built up some more tools (see ex. 3.5.33). If you're prepared to accept the Cauchy–Schwarz inequality for now, then the triangle inequality follows, because, by the properties of the inner product given in fact 2.1.1,

$$\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + 2 \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \leqslant \langle \mathbf{x}, \mathbf{x} \rangle + 2 |\langle \mathbf{x}, \mathbf{y} \rangle| + \langle \mathbf{y}, \mathbf{y} \rangle$$

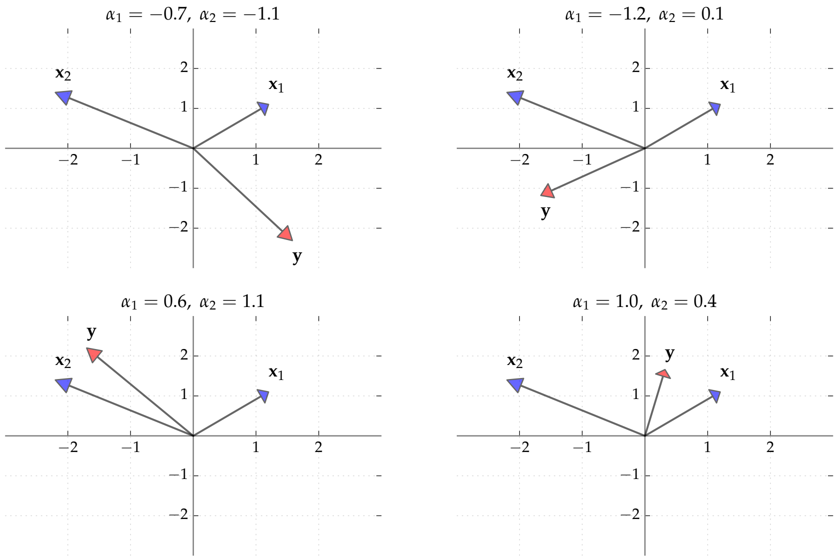


Figure 2.5 Linear combinations of $\mathbf{x}_1, \mathbf{x}_2$

Applying the Cauchy–Schwarz inequality leads to $\|\mathbf{x} + \mathbf{y}\|^2 \leqslant (\|\mathbf{x}\| + \|\mathbf{y}\|)^2$. Taking the square root gives the triangle inequality.

Given two vectors $\mathbf{x}$ and $\mathbf{y}$, the value $\|\mathbf{x} - \mathbf{y}\|$ has the interpretation of being the "distance" between these points. To see why, consult figure 2.4 again.

2.1.2 Linear Combinations and Span

One of the most elementary ways to work with vectors is to combine them using linear operations. Given vectors $\mathbf{x}_1, \ldots, \mathbf{x}_K$ in $\mathbb{R}^N$, a **linear combination** of these vectors is a new vector of the form

$$\mathbf{y} = \sum_{k=1}^{K} \alpha_k \mathbf{x}_k = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_K \mathbf{x}_K \tag{2.3}$$

for some collection of scalars $\alpha_1, \ldots, \alpha_K$ (i.e., with $\alpha_k \in \mathbb{R}$ for all k). Figure 2.5 shows four different linear combinations $\mathbf{y} = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2$ where $\mathbf{x}_1, \mathbf{x}_2$ are fixed vectors in $\mathbb{R}^2$ and the scalars α_1 and α_2 are varied.

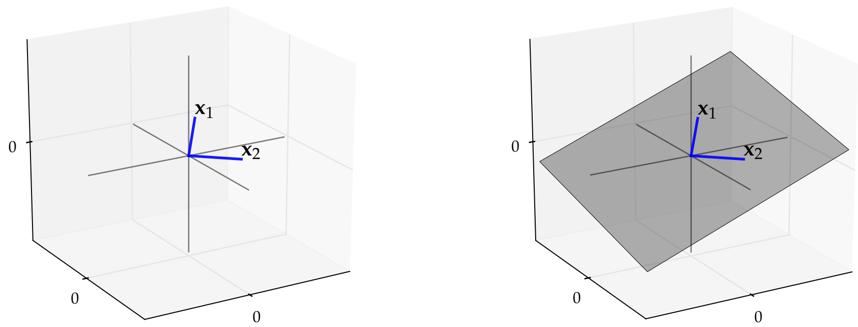


Figure 2.6 Span of $X = \{\mathbf{x}_1, \mathbf{x}_2\}$

Given any nonempty $X \subset \mathbb{R}^N$, the set of all vectors that can be made by (finite) linear combinations of elements of X is called the **span** of X, and denoted by span X. For example, the set of all linear combinations of $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ is

$$\operatorname{span} X := \left\{ \text{all vectors } \sum_{k=1}^{K} \alpha_k \mathbf{x}_k \text{ such that } \boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K \right\}$$

As will be discussed below, the span of certain collections of vectors turns out to have an intimate connection with existence of solutions to linear equations.

Example 2.1.1 By construction, the four vectors labeled $\mathbf{y}$ in figure 2.5 lie in the span of $X = \{\mathbf{x}_1, \mathbf{x}_2\}$. Looking at this picture might lead you to wonder whether *any* vector in $\mathbb{R}^2$ could be created as a linear combination of $\mathbf{x}_1, \mathbf{x}_2$. The answer is affirmative. We'll prove this in §2.1.5.

Example 2.1.2 Let $X = \{\mathbf{1}\} = \{(1, 1)\} \subset \mathbb{R}^2$. The span of X is all vectors of the form $\alpha\mathbf{1} = (\alpha, \alpha)$ with $\alpha \in \mathbb{R}$. This constitutes a line in the plane. Since we can take $\alpha = 0$, it follows that the origin $\mathbf{0}$ is in span X. In fact span X is the unique line in the plane that passes through both $\mathbf{0}$ and the vector $\mathbf{1} = (1, 1)$.

Example 2.1.3 Let $\mathbf{x}_1 = (3, 4, 2)$ and let $\mathbf{x}_2 = (3, -4, 0.4)$. The span of $\{\mathbf{x}_1, \mathbf{x}_2\}$ is a plane in $\mathbb{R}^3$ that passes through both of these vectors and the origin, as shown in figure 2.6.

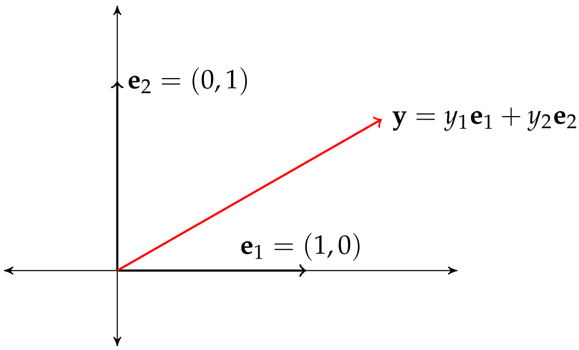


Figure 2.7 Canonical basis vectors in $\mathbb{R}^2$

Example 2.1.4 Consider the vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \subset \mathbb{R}^N$, where $\mathbf{e}_n$ has all zeros except for a 1 as the nth element:

$$\mathbf{e}_1 := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 := \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots, \quad \mathbf{e}_N := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

The case of $\mathbb{R}^2$ is illustrated in figure 2.7. The vectors $\mathbf{e}_1, \ldots, \mathbf{e}_N$ are called the **canonical basis vectors** of $\mathbb{R}^N$. One reason is that $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ spans all of $\mathbb{R}^N$. Here's a proof for $N = 2$: Observe that for any $\mathbf{y} \in \mathbb{R}^2$, we have

$$\mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ y_1 \end{pmatrix} = y_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + y_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = y_1\mathbf{e}_1 + y_2\mathbf{e}_2$$

Thus $\mathbf{y} \in \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2\}$. Since $\mathbf{y}$ is an arbitrary vector in $\mathbb{R}^2$, we have shown that $\{\mathbf{e}_1, \mathbf{e}_2\}$ spans $\mathbb{R}^2$.

Example 2.1.5 Consider the set $P := \{(x_1, x_2, 0) \in \mathbb{R}^3 : x_1, x_2 \in \mathbb{R}\}$. Graphically, P corresponds to the flat plane in $\mathbb{R}^3$, where the height coordinate is always zero. If we take $\mathbf{e}_1 = (1, 0, 0)$ and $\mathbf{e}_2 = (0, 1, 0)$, then given $\mathbf{y} = (y_1, y_2, 0) \in P$, we have $\mathbf{y} = y_1\mathbf{e}_1 + y_2\mathbf{e}_2$. In other words, any $\mathbf{y} \in P$ can be expressed as a linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2$. Equivalently, $P \subset \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2\}$.

The next fact follows directly from the definition of span.

Fact 2.1.3 If X, Y are nonempty subsets of $\mathbb{R}^N$ and $X \subset Y$, then $\operatorname{span} X \subset \operatorname{span} Y$.

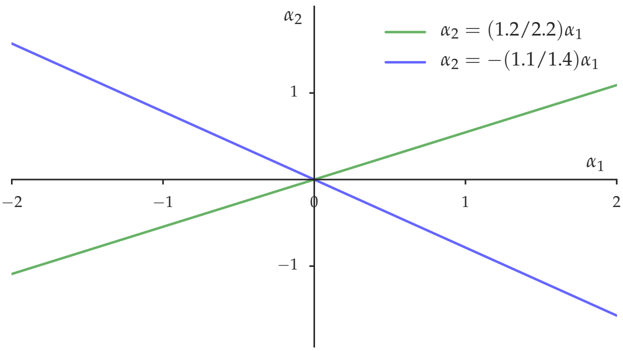


Figure 2.8 The only solution is $\alpha_1 = \alpha_2 = 0$

2.1.3 Linear Independence

Linear independence is an apparently simple concept with implications that stretch deep into many aspects of analysis. If you wish to understand when a matrix is invertible, or when a system of linear equations has a unique solution, or when a least squares estimate is uniquely defined, the most important foundational idea is linear independence of vectors.

Let's begin with the definition. Consider a set of vectors $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$. We can surely realize the origin $\mathbf{0}$ as a linear combination of these vectors, just by setting all of the scalars α_k in $\sum_{k=1}^K \alpha_k \mathbf{x}_k$ to zero. The set X is called linearly independent when this is the only possibility. That is, $X \subset \mathbb{R}^N$ is called **linearly independent** if

$$\alpha_1 \mathbf{x}_1 + \cdots + \alpha_K \mathbf{x}_K = \mathbf{0} \quad \Longrightarrow \quad \alpha_1 = \cdots = \alpha_K = 0 \tag{2.4}$$

We call X **linearly dependent** if it is not linearly independent.

Example 2.1.6 In figure 2.5 on page 14, the two vectors are $\mathbf{x}_1 = (1.2, 1.1)$ and $\mathbf{x}_2 = (-2.2, 1.4)$. Suppose that α_1 and α_2 are scalars with

$$\alpha_1 \begin{pmatrix} 1.2 \\ 1.1 \end{pmatrix} + \alpha_2 \begin{pmatrix} -2.2 \\ 1.4 \end{pmatrix} = \mathbf{0}$$

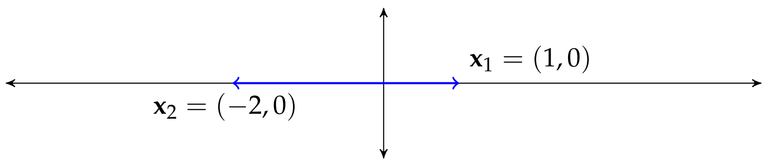
This translates to the linear, two-equation system shown in figure 2.8, where the unknowns are α_1 and α_2. The only solution is $\alpha_1 = \alpha_2 = 0$. Hence $\{\mathbf{x}_1, \mathbf{x}_2\}$ is linearly independent.

Example 2.1.7 The set of N canonical basis vectors $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$ is linearly independent in $\mathbb{R}^N$. To see this, let $\alpha_1, \dots, \alpha_N$ be coefficients such that $\sum_{n=1}^N \alpha_n \mathbf{e}_n = \mathbf{0}$. Equivalently,

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \alpha_N \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

In particular, $\alpha_n = 0$ for all n.

Example 2.1.8 Consider the vectors $\{\mathbf{x}_1, \mathbf{x}_2\}$ as given in



This pair fails to be linearly independent, since $\mathbf{x}_2 = -2\mathbf{x}_1$, and hence $2\mathbf{x}_1 + \mathbf{x}_2 = \mathbf{0}$.

How can we interpret linear independence? One way to understand it is as an indicator of the algebraic diversity of a given collection of vectors. In particular, in a linearly independent set, the span is relatively large, in the sense that every vector contributes to the span. Here's a formal statement of this idea.

Theorem 2.1.1 *Let $X := \{\mathbf{x}_1, \dots, \mathbf{x}_K\} \subset \mathbb{R}^N$. For $K > 1$, the following statements are equivalent:*

(i) X is linearly independent.

(ii) X_0 is a proper subset[2] of X $\implies$ $\operatorname{span} X_0$ is a proper subset of $\operatorname{span} X$.

(iii) No vector in X can be written as a linear combination of the others.

Exercise 2.4.15 asks you to check these equivalences. For now let's just step through them in the context of two examples. First consider the pair of canonical basis vectors $\{\mathbf{e}_1, \mathbf{e}_2\}$ in $\mathbb{R}^2$, as depicted in figure 2.7. As we saw in examples 2.1.4 and 2.1.7, this pair is linearly independent, and its span is all of $\mathbb{R}^2$. Both vectors contribute to the span, since removing either one reduces the span to just a line in $\mathbb{R}^2$. (For example, the span of $\{\mathbf{e}_1\}$ is just the horizontal axis in $\mathbb{R}^2$.) Neither one of this pair can be written as a linear combination of the other.

2. A is a proper subset of B if $A \subset B$ and $A \neq B$.

Next, consider instead the pair $\{\mathbf{x}_1, \mathbf{x}_2\}$ in example 2.1.8. These vectors fail to be linearly independent, as shown in that example. It is also clear that dropping either one does not change the span—it remains the horizontal axis in any case. Finally, we saw in example 2.1.8 that $\mathbf{x}_2 = -2\mathbf{x}_1$, which means that each vector can be written as a linear combination of the other.

Fact 2.1.4 If $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ is linearly independent, then

(i) every subset of X is linearly independent,

(ii) X does not contain $\mathbf{0}$, and

(iii) $X \cup \{\mathbf{x}\}$ is linearly independent for all $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{x} \notin \operatorname{span} X$.

The proof is a solved exercise (ex. 2.4.16 on page 36).

2.1.3.1 Linear Independence and Uniqueness

As we'll see below, the problem of existence of solutions to systems of linear equations comes down to whether or not a given point is contained in the span of a collection of vectors (which typically correspond to the columns of a matrix). This depends in general on the size of the span, and the size of the span in turn depends on whether or not the vectors are linearly independent.

Given that linear independence is the key condition for existence of solutions, it's surprising at first to learn that linear independence is the key condition for uniqueness as well. As we'll see, the connection between linear independence and uniqueness stems from the following result.

Theorem 2.1.2 *Let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ be any collection of vectors in $\mathbb{R}^N$. The following statements are equivalent:*

(i) X is linearly independent.

(ii) For each $\mathbf{y} \in \mathbb{R}^N$ there exists at most one set of scalars $\alpha_1, \ldots, \alpha_K$ such that

$$\mathbf{y} = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_K \mathbf{x}_K \tag{2.5}$$

Proof. ((i) $\implies$ (ii)) Let X be linearly independent and pick any $\mathbf{y} \in \mathbb{R}^N$. Suppose that there are two sets of scalars such that (2.5) holds. In particular, suppose that $\mathbf{y} = \sum_{k=1}^K \alpha_k \mathbf{x}_k = \sum_{k=1}^K \beta_k \mathbf{x}_k$. It follows from the second equality that $\sum_{k=1}^K (\alpha_k - \beta_k)\mathbf{x}_k = \mathbf{0}$. By linear independence, we then have $\alpha_k = \beta_k$ for all k. In other words, the representation is unique.

((ii) $\implies$ (i)) If (ii) holds, then there exists at most one set of scalars such that $\mathbf{0} = \sum_{k=1}^K \alpha_k \mathbf{x}_k$. Because $\alpha_1 = \cdots = \alpha_k = 0$ has this property, we conclude that no nonzero scalars yield $\mathbf{0} = \sum_{k=1}^K \alpha_k \mathbf{x}_k$. In other words, X is linearly independent. $\square$

2.1.4 Linear Subspaces

We will find it rewarding to study more closely the structure of spans generated by a collection of vectors. One of the defining features of the span of a set X is that it is "closed" under the linear operations of vector addition and scalar multiplication, in the sense that

(i) $\mathbf{x}, \mathbf{y} \in \operatorname{span} X \implies \mathbf{x} + \mathbf{y} \in \operatorname{span} X$, and

(ii) $\mathbf{y} \in \operatorname{span} X$ and $\gamma \in \mathbb{R} \implies \gamma\mathbf{y} \in \operatorname{span} X$.

For example, (i) holds because the sum of two linear combinations of elements of X is another linear combination of elements of X.

The notion of a set being closed under scalar multiplication and vector addition is important enough to have its own name: A nonempty subset S of $\mathbb{R}^N$ is called a **linear subspace** (or just **subspace**) of $\mathbb{R}^N$ if

$$\mathbf{x}, \mathbf{y} \in S \text{ and } \alpha, \beta \in \mathbb{R} \implies \alpha\mathbf{x} + \beta\mathbf{y} \in S \tag{2.6}$$

Example 2.1.9 It follows from the preceding discussion that if X is any nonempty subset of $\mathbb{R}^N$, then $\operatorname{span} X$ is a linear subspace of $\mathbb{R}^N$. For this reason, $\operatorname{span} X$ is often called the **linear subspace spanned by X.**

Example 2.1.10 Given any $\mathbf{a} \in \mathbb{R}^N$, the set $A := \{\mathbf{x} \in \mathbb{R}^N : \langle \mathbf{a}, \mathbf{x} \rangle = 0\}$ is a linear subspace of $\mathbb{R}^N$. To see this let $\mathbf{x}, \mathbf{y} \in A$ and let $\alpha, \beta \in \mathbb{R}$. We claim that $\mathbf{z} := \alpha\mathbf{x} + \beta\mathbf{y} \in A$, or, equivalently, that $\langle \mathbf{a}, \mathbf{z} \rangle = 0$. This is true because

$$\langle \mathbf{a}, \mathbf{z} \rangle = \langle \mathbf{a}, \alpha\mathbf{x} + \beta\mathbf{y} \rangle = \alpha \langle \mathbf{a}, \mathbf{x} \rangle + \beta \langle \mathbf{a}, \mathbf{y} \rangle = 0 + 0 = 0$$

Example 2.1.11 The entire space $\mathbb{R}^N$ is a linear subspace of $\mathbb{R}^N$ because any linear combination of N-vectors is an N-vector.

To visualize subspaces in $\mathbb{R}^3$, think of lines and planes that pass through the origin. Here are some elementary facts about linear subspaces:

Fact 2.1.5 If S is a linear subspace of $\mathbb{R}^N$, then

(i) $\mathbf{0} \in S$,

(ii) $X \subset S \implies \operatorname{span} X \subset S$, and

(iii) $\operatorname{span} S = S$.

There's also one deep result about linear subspaces we need to cover, which forms a cornerstone of many foundational results:

Theorem 2.1.3 *Let S be a linear subspace of $\mathbb{R}^N$. If S is spanned by K vectors, then any linearly independent subset of S has at most K vectors.*

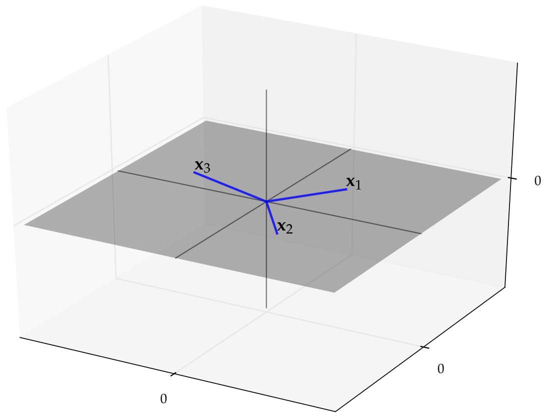


Figure 2.9 Any three vectors in P are linearly dependent

In other words, if there exists a set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ with $S \subset \operatorname{span} X$, then any subset of S with more than K vectors will be linearly dependent. The proof can be found in most texts on linear algebra (e.g., §3.5 of Jänich 1994).

Example 2.1.12 We saw in example 2.1.4 that $\mathbb{R}^2$ is spanned by the pair $\{\mathbf{e}_1, \mathbf{e}_2\}$, where $\mathbf{e}_i$ is the ith canonical basis vector in $\mathbb{R}^2$. (See also figure 2.7.) It follows immediately from this fact and theorem 2.1.3 that the three vectors in $\mathbb{R}^2$ shown in figure 2.1 are linearly dependent.

Example 2.1.13 Consider the plane $P := \{(x_1, x_2, 0) \in \mathbb{R}^3 : x_1, x_2 \in \mathbb{R}\}$ from example 2.1.5. We saw in that example that P can be spanned by two vectors. As a consequence of theorem 2.1.3, we now know that any three vectors in this plane—such as the three shown in figure 2.9—are linearly dependent.

2.1.5 Bases and Dimension

Consider again the pair $\mathbf{x}_1, \mathbf{x}_2$ shown in figure 2.5 on page 14, and the four different vectors labeled $\mathbf{y}$ that we created from $\mathbf{x}_1, \mathbf{x}_2$ by way of linear combinations. Eyeballing the figure gives the impression that any $\mathbf{y} \in \mathbb{R}^2$ could be constructed as a linear combination of $\mathbf{x}_1, \mathbf{x}_2$ with suitable choice of the scalars α_1, α_2. Indeed this is true. The reason is that the pair $\{\mathbf{x}_1, \mathbf{x}_2\}$ is linearly independent (see example 2.1.6

on page 17), and any pair of linearly independent vectors in $\mathbb{R}^2$ spans $\mathbb{R}^2$. Here's a statement of this result for the general case:

Theorem 2.1.4 *Let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be any N vectors in $\mathbb{R}^N$. The following statements are equivalent:*

(i) $\operatorname{span} X = \mathbb{R}^N$.

(ii) X is linearly independent.

Proof. ((i) $\implies$ (ii)) Suppose that $\operatorname{span} X = \mathbb{R}^N$ but X is not linearly independent. Then, by theorem 2.1.1, there exists a proper subset X_0 of X with $\operatorname{span} X_0 = \operatorname{span} X$. Since X_0 is a proper subset of X it contains $K < N$ elements. We now have K vectors spanning $\mathbb{R}^N$. In particular, the span of K vectors contains the $N > K$ linearly independent vectors $\mathbf{e}_1, \ldots, \mathbf{e}_N$. This contradicts theorem 2.1.3.

((ii) $\implies$ (i)) Suppose that X is linearly independent and yet there exists an $\mathbf{x} \in \mathbb{R}^N$ with $\mathbf{x} \notin \operatorname{span} X$. By fact 2.1.4, it follows that the $N+1$ element set $X \cup \{\mathbf{x}\} \subset \mathbb{R}^N$ is linearly independent. Since $\mathbb{R}^N$ is spanned by the N canonical basis vectors, this statement also contradicts theorem 2.1.3. □

We now come to a key definition. Let S be a linear subspace of $\mathbb{R}^N$ and let $B \subset S$. The set B is called a **basis** of S if

(i) B spans S and

(ii) B is linearly independent.

The plural of basis is **bases**. In view of theorem 2.1.2, when B is a basis of S, each point in S has exactly one representation as a linear combination of elements of B.

It follows immediately from theorem 2.1.4 that any N linearly independent vectors in $\mathbb{R}^N$ form a basis of $\mathbb{R}^N$.

Example 2.1.14 The set of canonical basis vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \subset \mathbb{R}^N$ described in example 2.1.4 is linearly independent and spans all of $\mathbb{R}^N$. As a result, it provides a basis for $\mathbb{R}^N$—as anticipated by the name.

Example 2.1.15 The pair $\{\mathbf{x}_1, \mathbf{x}_2\}$ from figure 2.5 (page 14) forms a basis of $\mathbb{R}^2$.

Example 2.1.16 The pair $\{\mathbf{e}_1, \mathbf{e}_2\}$ is a basis for the set P defined in example 2.1.5.

Here are the two most fundamental results about bases:

Theorem 2.1.5 *If S is a linear subspace of $\mathbb{R}^N$ distinct from $\{\mathbf{0}\}$, then*

(i) S has at least one basis and

(ii) every basis of S has the same number of elements.

The proof of part (i) is not particularly hard. See, for example, section 3.2 of Jänich (1994). Part (ii) follows from theorem 2.1.3 and is left as an exercise (ex. 2.4.17).

If S is a linear subspace of $\mathbb{R}^N$, then the common number identified in theorem 2.1.5 is called the **dimension** of S, and written as $\dim S$.

Example 2.1.17 For $P := \{(x_1, x_2, 0) \in \mathbb{R}^3 : x_1, x_2 \in \mathbb{R}\}$ we have $\dim P = 2$ because $\{\mathbf{e}_1, \mathbf{e}_2\} \subset \mathbb{R}^3$ is a basis (see example 2.1.5) and $\{\mathbf{e}_1, \mathbf{e}_2\}$ has two elements.

Example 2.1.18 $\dim \mathbb{R}^N = N$ because $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \subset \mathbb{R}^N$ is a basis.

Example 2.1.19 A line $\{\alpha \mathbf{x} \in \mathbb{R}^N : \alpha \in \mathbb{R}\}$ through the origin is one dimensional.

In $\mathbb{R}^N$ the singleton subspace $\{\mathbf{0}\}$ is said to have zero dimension.

If we take a set of K vectors, then how large will its span be in terms of dimension? The next theorem answers this question.

Theorem 2.1.6 *Let* $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$. *Then*

(i) $\dim \operatorname{span} X \leqslant K$ *and*

(ii) $\dim \operatorname{span} X = K$ *if and only if* X *is linearly independent.*

Exercise 2.4.19 asks you to prove these results.

Let's finish this section with facts that can be deduced from the preceding results.

Fact 2.1.6 The following statements are true:

(i) Let S and S' be K-dimensional linear subspaces of $\mathbb{R}^N$. If $S \subset S'$, then $S = S'$.

(ii) If S is an M-dimensional linear subspace of $\mathbb{R}^N$ and $M < N$, then $S \neq \mathbb{R}^N$.

Part (i) of fact 2.1.6 implies that the only N-dimensional linear subspace of $\mathbb{R}^N$ is $\mathbb{R}^N$ itself.

2.1.6 Linear Maps

The single most important class of functions in applied mathematics is the linear functions. In high school we are told that linear functions as those whose graph is a straight line. Here's a better definition: A function $T \colon \mathbb{R}^K \to \mathbb{R}^N$ is **linear** if

$$T(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha T \mathbf{x} + \beta T \mathbf{y} \quad \text{for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^K \text{ and } \alpha, \beta \in \mathbb{R} \tag{2.7}$$

(Following a common convention, we'll write linear functions with uppercase letters and omit the parenthesis around the argument where no confusion arises. This convention has come about because the action of linear maps is essentially isomorphic to multiplication of vectors by matrices. More on that topic soon.)

Example 2.1.20 The function $T \colon \mathbb{R} \to \mathbb{R}$ defined by $Tx = 2x$ is linear because, for any α, β, x, y in $\mathbb{R}$, we have $T(\alpha x + \beta y) = 2(\alpha x + \beta y) = \alpha 2x + \beta 2y = \alpha Tx + \beta Ty$.

Example 2.1.21 Given $\mathbf{a} \in \mathbb{R}^K$, the function $T \colon \mathbb{R}^K \to \mathbb{R}$ defined by $T\mathbf{x} = \langle \mathbf{a}, \mathbf{x} \rangle$ is linear. Indeed, by the rules for inner products on page 13, for any α, β in $\mathbb{R}$ and $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^K$ we have

$$T(\alpha \mathbf{x} + \beta \mathbf{y}) = \langle \mathbf{a}, \alpha \mathbf{x} + \beta \mathbf{y} \rangle = \alpha \langle \mathbf{a}, \mathbf{x} \rangle + \beta \langle \mathbf{a}, \mathbf{y} \rangle = \alpha T\mathbf{x} + \beta T\mathbf{y}$$

Example 2.1.22 The function $f \colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ fails to be linear because, if $\alpha = \beta = x = y = 1$, then $f(\alpha x + \beta y) = 4$, while $\alpha f(x) + \beta f(y) = 2$.

Example 2.1.23 The function $f \colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 1 + 2x$ is not linear because if $\alpha = \beta = x = y = 1$, then $f(\alpha x + \beta y) = f(2) = 5$, while $\alpha f(x) + \beta f(y) = 3 + 3 = 6$. This kind of function is called an **affine** function. We see that identifying linear functions with functions whose graph is a straight line is not correct.

The definition in (2.7) tells us directly that if T is linear then the exchange of order in $T[\sum_{k=1}^K \alpha_k \mathbf{x}_k] = \sum_{k=1}^K \alpha_k T\mathbf{x}_k$ will be valid whenever $K = 2$. A simple inductive argument extends this to arbitrary K. As an application of this fact, consider the following: As discussed in example 2.1.4, any $\mathbf{x} \in \mathbb{R}^K$ can be expressed in terms of the basis vectors as $\sum_{k=1}^K \alpha_k \mathbf{e}_k$, for some suitable choice of scalars. Hence, for a linear function T, its range, denoted $\operatorname{rng} T$, is the set of all points of the form

$$T\mathbf{x} = T\left[\sum_{k=1}^K \alpha_k \mathbf{e}_k\right] = \sum_{k=1}^K \alpha_k T\mathbf{e}_k$$

as we vary $\alpha_1, \ldots, \alpha_K$ over all scalar combinations. (See §15.2 for the definition of range.) In other words, the range of a linear map is the span of the image of the canonical basis functions. This will prove to be important later on. The next fact summarizes.

Fact 2.1.7 If $T \colon \mathbb{R}^K \to \mathbb{R}^N$ is a linear map, then

$$\operatorname{rng} T = \operatorname{span} V, \quad \text{where } V := \{T\mathbf{e}_1, \ldots, T\mathbf{e}_K\}$$

Soon we'll turn to the topic of determining when linear functions are bijections, an issue that is intimately related to invertibility of matrices. To this end it's useful to note that, for linear functions, the property of being one-to-one can be determined by examining the set of points it maps to the origin. To express this idea, for any $T \colon \mathbb{R}^K \to \mathbb{R}^N$, we define the **null space** or **kernel** of T as

$$\operatorname{null} T := \{\mathbf{x} \in \mathbb{R}^K : T\mathbf{x} = \mathbf{0}\}$$

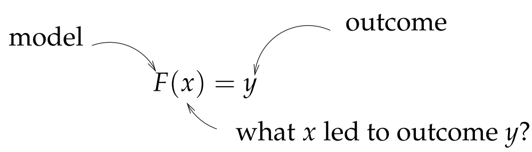


Figure 2.10 Inverse problem

Fact 2.1.8 If $T\colon \mathbb{R}^K \to \mathbb{R}^N$ is linear, then

(i) $\operatorname{null} T$ is a linear subspace of $\mathbb{R}^K$ and

(ii) $\operatorname{null} T = \{\mathbf{0}\}$ if and only if T is one-to-one.

The proofs are straightforward. For example, if $T\mathbf{x} = T\mathbf{y}$ for some $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$, then $T(\mathbf{x} - \mathbf{y}) = \mathbf{0}$, and hence $\mathbf{x} - \mathbf{y} \in \operatorname{null} T$. So if $\operatorname{null} T = \{\mathbf{0}\}$, then $T\mathbf{x} = T\mathbf{y}$ implies $\mathbf{x} = \mathbf{y}$, which means that T is one-to-one.

2.1.7 Linear Independence and Bijections

Many statistical problems are “inverse" problems, in the sense that we observe outcomes and wish to determine what generated them. For example, we might want to know what consumer preferences led to observed market behavior, or what kinds of expectations led to a given shift in exchange rates.

Consider the generic inverse problem in figure 2.10, where F and y are given, and we seek to obtain the unknown object x. Two immediate questions are: Does this problem have a solution? and Is it unique? To provide general answers to these questions, we need to know whether F is one-to-one, onto, etc. (see §15.2 for definitions and further discussion). The best case is when F is a bijection, for then we know that a unique solution x exists for every possible y.

In general, functions can be onto, one-to-one, bijections, or none of the above. However, for linear functions from $\mathbb{R}^N$ to $\mathbb{R}^N$, the first three properties are all equivalent! The next theorem gives details.

Theorem 2.1.7 *If T is a linear function from $\mathbb{R}^N$ to $\mathbb{R}^N$, then all of the following are equivalent:*

(i) T is a bijection.

(ii) T is onto.

(iii) T is one-to-one.

(iv) $\operatorname{null} T = \{\mathbf{0}\}$.

(v) $V := \{T\mathbf{e}_1, \ldots, T\mathbf{e}_N\}$ is linearly independent.

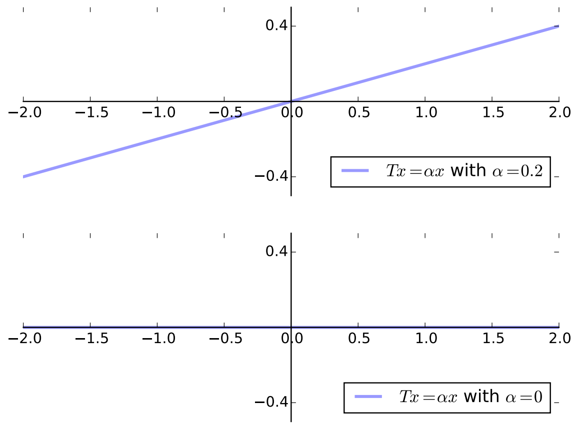


Figure 2.11 The case of $N = 1$, nonsingular and singular

(vi) $V := \{T\mathbf{e}_1, \ldots, T\mathbf{e}_N\}$ *forms a basis of* $\mathbb{R}^N$.

If any one of these conditions is true, then T is called **nonsingular**. (Equivalently, a nonsingular function is a linear bijection.) Otherwise T is called **singular**. The proof of theorem 2.1.7 is a solved exercise (ex. 2.4.21). Figure 2.11 provides intuition for the case of $N = 1$. In the top panel all conditions in theorem 2.1.7 are satisfied. In the lower panel none are. In particular, we can see that the condition for T to be one-to-one and onto is exactly the same (i.e., $\alpha \neq 0$).

If T is nonsingular, then, being a bijection, it must have an inverse function T^{-1} that is also a bijection (fact 15.2.1 on page 410). It turns out that this inverse function inherits the linearity of T (see ex. 2.4.20). In summary,

Fact 2.1.9 If $T\colon \mathbb{R}^N \to \mathbb{R}^N$ is nonsingular, then so is T^{-1}.

Theorem 2.1.7 only applies to linear maps between spaces of the *same* dimension. When linear functions map across distinct dimensions the situation changes:

Theorem 2.1.8 *For a linear map* T *from* $\mathbb{R}^K \to \mathbb{R}^N$*, the following statements are true:*

(i) If $K < N$*, then* T *is not onto.*

(ii) If $K > N$*, then* T *is not one-to-one.*

The most important implication is that if $N \neq K$, then we can forget about bijections. The proof of theorem 2.1.8 is a solved exercise (ex. 2.4.22).

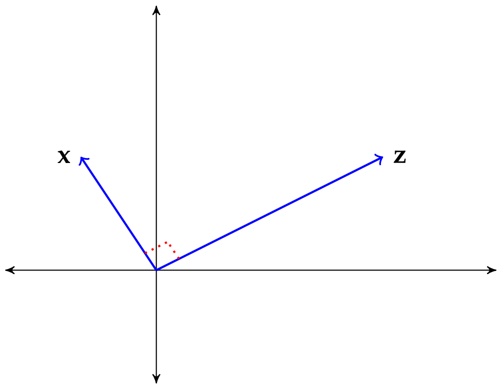


Figure 2.12 $\mathbf{x} \perp \mathbf{z}$

2.2 Orthogonality

One of the core concepts in this book is orthogonality, not just of vectors but also of more complex objects such as random variables. Let's begin with the vector definition and some key implications.

2.2.1 Definition and Basic Properties

Let $\mathbf{x}$ and $\mathbf{z}$ be vectors in $\mathbb{R}^N$. If $\langle \mathbf{x}, \mathbf{z} \rangle = 0$, then we write $\mathbf{x} \perp \mathbf{z}$ and call $\mathbf{x}$ and $\mathbf{z}$ **orthogonal**. In $\mathbb{R}^2$, $\mathbf{x}$ and $\mathbf{z}$ are orthogonal when they are perpendicular to one another, as in figure 2.12. For $\mathbf{x} \in \mathbb{R}^N$ and $S \subset \mathbb{R}^N$, we say that **$\mathbf{x}$ is orthogonal to S** if $\mathbf{x} \perp \mathbf{z}$ for all $\mathbf{z} \in S$ (figure 2.13), and we write $\mathbf{x} \perp S$. A set of vectors $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\} \subset \mathbb{R}^N$ is called an **orthogonal set** if its elements are mutually orthogonal, that is, if $\mathbf{z}_j \perp \mathbf{z}_k$ whenever j and k are distinct.

Fact 2.2.1 (Pythagorian law) If $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$ is an orthogonal set, then

$$\|\mathbf{z}_1 + \cdots + \mathbf{z}_K\|^2 = \|\mathbf{z}_1\|^2 + \cdots + \|\mathbf{z}_K\|^2$$

Orthogonal sets and linear independence are closely related. For example,

Fact 2.2.2 If $O \subset \mathbb{R}^N$ is an orthogonal set and $\mathbf{0} \notin O$, then O is linearly independent.

While not every linearly independent set is orthogonal, an important partial converse to fact 2.2.2 is given in §2.2.4.

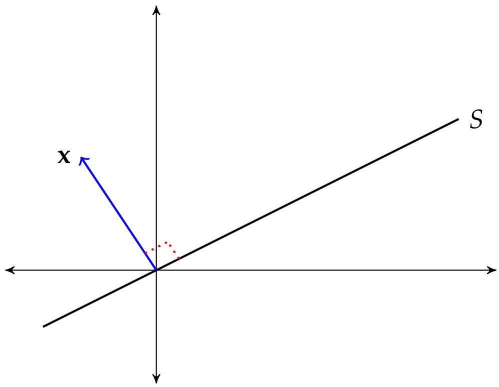


Figure 2.13 $\mathbf{x} \perp S$

An orthogonal set $O \subset \mathbb{R}^N$ is called an **orthonormal set** if $\|\mathbf{u}\| = 1$ for all $\mathbf{u} \in O$. An orthonormal set that lies in and spans a linear subspace S of $\mathbb{R}^N$ is called an **orthonormal basis** of S. It is, necessarily, a basis of S. (Why?) The standard example of an orthonormal basis for all of $\mathbb{R}^N$ is the canonical basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$.

By definition, if $O = \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ is any basis of S, then, for any $\mathbf{x} \in S$, we can find unique scalars $\alpha_1, \ldots, \alpha_K$ such that $\mathbf{x} = \sum_{k=1}^K \alpha_k \mathbf{u}_k$. While the values of these scalars are not always transparent, for an orthonormal basis they are easy to compute:

Fact 2.2.3 If $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ is an orthonormal set and $\mathbf{x} \in \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$, then

$$\mathbf{x} = \sum_{k=1}^{K} \langle \mathbf{x}, \mathbf{u}_k \rangle \, \mathbf{u}_k \tag{2.8}$$

The proof is an exercise. Given $S \subset \mathbb{R}^N$, the **orthogonal complement** of S is defined as

$$S^\perp := \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} \perp S\}$$

Figure 2.14 gives an example in $\mathbb{R}^2$.

Fact 2.2.4 For any nonempty $S \subset \mathbb{R}^N$, the set $S^\perp$ is a linear subspace of $\mathbb{R}^N$.

Indeed, if $\mathbf{x}, \mathbf{y} \in S^\perp$ and $\alpha, \beta \in \mathbb{R}$, then $\alpha\mathbf{x} + \beta\mathbf{y} \in S^\perp$ because, for any $\mathbf{z} \in S$,

$$\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle = \alpha \times 0 + \beta \times 0 = 0$$

Fact 2.2.5 For $S \subset \mathbb{R}^N$, we have $S \cap S^\perp = \{\mathbf{0}\}$.

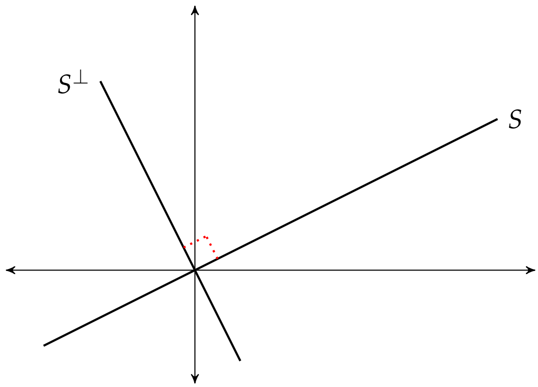


Figure 2.14 Orthogonal complement of S

2.2.2 The Orthogonal Projection Theorem

A central problem in linear regression and many other applications is approximation of some $\mathbf{y}$ of $\mathbb{R}^N$ by an element of a given subspace S of $\mathbb{R}^N$. Stated more precisely, the problem is, given $\mathbf{y}$ and S, to find the closest element $\hat{\mathbf{y}}$ of S to $\mathbf{y}$. Closeness is in terms of Euclidean norm, so $\hat{\mathbf{y}}$ is the minimizer of $\|\mathbf{y} - \mathbf{z}\|$ over all $\mathbf{z} \in S$:

$$\hat{\mathbf{y}} = \operatorname*{argmin}_{\mathbf{z} \in S} \|\mathbf{y} - \mathbf{z}\| \tag{2.9}$$

The next theorem tells us that a solution $\hat{\mathbf{y}}$ to this minimization problem always exists, as well as providing a means to identify it.

Theorem 2.2.1 (Orthogonal Projection Theorem I) *Let $\mathbf{y} \in \mathbb{R}^N$ and let S be any nonempty linear subspace of $\mathbb{R}^N$. The following statements are true:*

(i) The optimization problem (2.9) has exactly one solution.

(ii) $\hat{\mathbf{y}} \in \mathbb{R}^N$ solves (2.9) if and only if $\hat{\mathbf{y}} \in S$ and $\mathbf{y} - \hat{\mathbf{y}} \perp S$.

The unique solution $\hat{\mathbf{y}}$ is called the **orthogonal projection of y onto** S.

The intuition is easy to grasp from a graphical presentation. Figure 2.15 illustrates. Looking at the figure, we can see that the closest point to $\mathbf{y}$ in S is the one point $\hat{\mathbf{y}} \in S$ such that $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to S.

For a full proof see, for example, theorem 5.16 of Çinlar and Vanderbei (2013). Let's just cover sufficiency of the conditions in part (ii): Let $\mathbf{y} \in \mathbb{R}^N$ and let S be a linear subspace of $\mathbb{R}^N$. Let $\hat{\mathbf{y}}$ be a vector in S satisfying $\mathbf{y} - \hat{\mathbf{y}} \perp S$. Let $\mathbf{z}$ be any other point

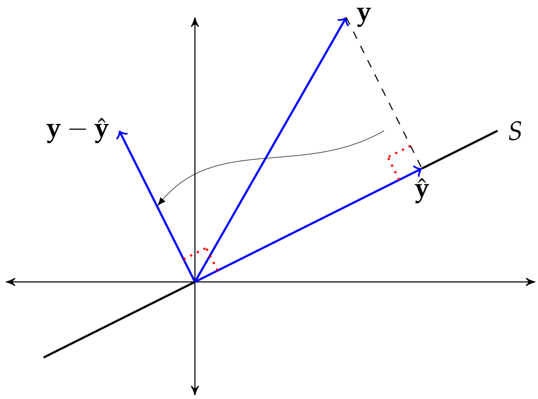


Figure 2.15 Orthogonal projection

in S. We have

$$\|\mathbf{y}-\mathbf{z}\|^2 = \|(\mathbf{y}-\hat{\mathbf{y}}) + (\hat{\mathbf{y}}-\mathbf{z})\|^2 = \|\mathbf{y}-\hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}}-\mathbf{z}\|^2$$

The second equality follows from $\mathbf{y}-\hat{\mathbf{y}} \perp S$ (why?) and the Pythagorian law. Since $\mathbf{z}$ was an arbitrary point in S, we have $\|\mathbf{y}-\mathbf{z}\| \geqslant \|\mathbf{y}-\hat{\mathbf{y}}\|$ for all $\mathbf{z} \in S$. Hence (2.9) holds.

Example 2.2.1 Let $\mathbf{y} \in \mathbb{R}^N$ and let $\mathbf{1} \in \mathbb{R}^N$ be the vector of ones. Let S be the set of constant vectors in $\mathbb{R}^N$, meaning that all elements are equal. Evidently S is the span of $\{\mathbf{1}\}$. The orthogonal projection of $\mathbf{y}$ onto S is $\hat{\mathbf{y}} := \bar{y}\mathbf{1}$, where $\bar{y} := \frac{1}{N}\sum_{n=1}^N y_n$. To see this, note that $\hat{\mathbf{y}} \in S$ clearly holds. Hence we only need to check that $\mathbf{y}-\hat{\mathbf{y}}$ is orthogonal to S, for which it suffices to show that $\langle \mathbf{y}-\hat{\mathbf{y}}, \mathbf{1}\rangle = 0$ (see ex. 2.4.14 on page 36). This is true because

$$\langle \mathbf{y}-\hat{\mathbf{y}}, \mathbf{1}\rangle = \langle \mathbf{y}, \mathbf{1}\rangle - \langle \hat{\mathbf{y}}, \mathbf{1}\rangle = \sum_{n=1}^N y_n - \bar{y}\,\langle \mathbf{1}, \mathbf{1}\rangle = 0$$

2.2.3 Projection as a Mapping

In view of theorem 2.2.1, for each fixed linear subspace S in $\mathbb{R}^N$, the operation

$$\mathbf{y} \;\mapsto\; \text{the orthogonal projection of } \mathbf{y} \text{ onto } S$$

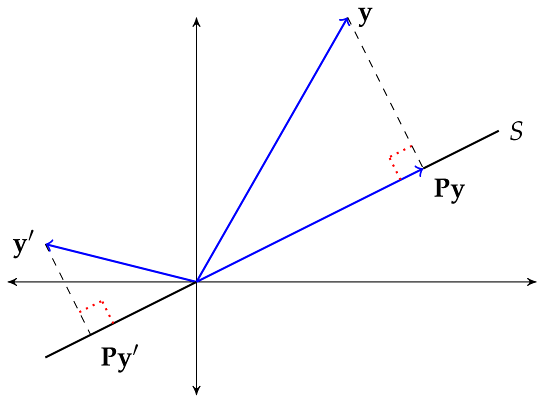


Figure 2.16 Orthogonal projection under $\mathbf{P}$

is a well-defined function from $\mathbb{R}^N$ to $\mathbb{R}^N$. The function is typically denoted by $\mathbf{P}$. For each $\mathbf{y} \in \mathbb{R}^N$, the symbol $\mathbf{P}\mathbf{y}$ represents the image of $\mathbf{y}$ under $\mathbf{P}$, which is the orthogonal projection $\hat{\mathbf{y}}$. $\mathbf{P}$ is called the **orthogonal projection onto** S, and we write

$$\mathbf{P} = \operatorname{proj} S$$

Figure 2.16 illustrates the action of $\mathbf{P}$ on two different vectors.

Using this notation, we can restate the orthogonal projection theorem, as well as adding some properties of $\mathbf{P}$:

Theorem 2.2.2 (Orthogonal Projection Theorem II) *Let S be any linear subspace of $\mathbb{R}^N$, and let $\mathbf{P} = \operatorname{proj} S$. The following statements are true:*

(i) $\mathbf{P}$ *is a linear function.*

Moreover, for any $\mathbf{y} \in \mathbb{R}^N$, we have

(ii) $\mathbf{P}\mathbf{y} \in S$,

(iii) $\mathbf{y} - \mathbf{P}\mathbf{y} \perp S$,

(iv) $\|\mathbf{y}\|^2 = \|\mathbf{P}\mathbf{y}\|^2 + \|\mathbf{y} - \mathbf{P}\mathbf{y}\|^2$,

(v) $\|\mathbf{P}\mathbf{y}\| \leqslant \|\mathbf{y}\|$,

(vi) $\mathbf{P}\mathbf{y} = \mathbf{y}$ *if and only if* $\mathbf{y} \in S$, *and*

(vii) $\mathbf{P}\mathbf{y} = \mathbf{0}$ *if and only if* $\mathbf{y} \in S^{\perp}$.

These results are not difficult to prove, given theorem 2.2.1. Linearity of $\mathbf{P}$ is left as an exercise (ex. 2.4.29). Parts (ii)–(iii) follow directly from theorem 2.2.1. To see (iv), observe that $\mathbf{y}$ can be decomposed as $\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{y} - \mathbf{P}\mathbf{y}$. Now apply the Pythagorean law (page 27). Part (v) follows from part (iv). Part (vi) follows from the definition of $\mathbf{P}\mathbf{y}$ as the closest point to $\mathbf{y}$ in S. Part (vii) is an exercise.

Fact 2.2.6 If $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ is an orthonormal basis for S, then, for each $\mathbf{y} \in \mathbb{R}^N$,

$$\mathbf{P}\mathbf{y} = \sum_{k=1}^{K} \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k \tag{2.10}$$

Fact 2.2.6 is a fundamental result. It's true because the right-hand side of (2.10) clearly lies in S (being a linear combination of basis functions) and, for any $\mathbf{u}_j$ in the basis set

$$\left\langle \mathbf{y} - \mathbf{P}\mathbf{y}, \mathbf{u}_j \right\rangle = \left\langle \mathbf{y}, \mathbf{u}_j \right\rangle - \sum_{k=1}^{K} \langle \mathbf{y}, \mathbf{u}_k \rangle \left\langle \mathbf{u}_k, \mathbf{u}_j \right\rangle = \left\langle \mathbf{y}, \mathbf{u}_j \right\rangle - \left\langle \mathbf{y}, \mathbf{u}_j \right\rangle = 0$$

This is enough to confirm that $\mathbf{y} - \mathbf{P}\mathbf{y} \perp S$ (see ex. 2.4.14).

Example 2.2.2 Recall example 2.2.1, where we showed that the projection of $\mathbf{y} \in \mathbb{R}^N$ onto span$\{\mathbf{1}\}$ is $\bar{y}\mathbf{1}$, where $\bar{y}$ is the "sample mean" $\bar{y} := \frac{1}{N}\sum_{n=1}^{N} y_n$. We can see this from (2.10) too. To apply (2.10), we just need to find an orthonormal basis for span$\{\mathbf{1}\}$. The obvious candidate is $\{N^{-1/2}\mathbf{1}\}$. Applying (2.10) now gives $\mathbf{P}\mathbf{y} = \left\langle N^{-1/2}\mathbf{1}, \mathbf{y} \right\rangle N^{-1/2}\mathbf{1}$. As before, this leads us to $\bar{y}\mathbf{1}$.

There's one more essential property of $\mathbf{P}$ that we need to make note of: Suppose that we have two linear subspaces S_1 and S_2 of $\mathbb{R}^N$, where $S_1 \subset S_2$. What then is the difference between (1) first projecting a point onto the bigger subspace S_2, and then projecting the result onto the smaller subspace S_1, and (2) projecting directly to the smaller subspace S_1? The answer is none—we get the same result:

Fact 2.2.7 Let S_i be a linear subspace of $\mathbb{R}^N$ for $i = 1, 2$ and let $\mathbf{P}_i = \operatorname{proj} S_i$. If $S_1 \subset S_2$, then

$$\mathbf{P}_1\mathbf{P}_2\mathbf{y} = \mathbf{P}_2\mathbf{P}_1\mathbf{y} = \mathbf{P}_1\mathbf{y} \quad \text{for all } \mathbf{y} \in \mathbb{R}^N$$

2.2.4 The Residual Projection

Consider the setting of the orthogonal projection theorem. Our interest is in projecting $\mathbf{y}$ onto S, where S is a linear subspace of $\mathbb{R}^N$. The closest point to $\mathbf{y}$ in S is $\hat{\mathbf{y}} := \mathbf{P}\mathbf{y}$ where $\mathbf{P} = \operatorname{proj} S$. Unless $\mathbf{y}$ was already in S, some error $\mathbf{y} - \mathbf{P}\mathbf{y}$ remains. Tracking and managing this residual will be important to us, so let's introduce an operator $\mathbf{M}$ that takes $\mathbf{y} \in \mathbb{R}^N$ and returns the residual. We can define it as

$$\mathbf{M} := \mathbf{I} - \mathbf{P} \tag{2.11}$$

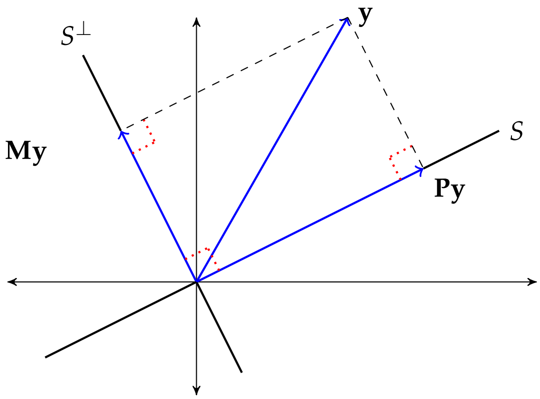


Figure 2.17 The residual projection

where $\mathbf{I}$ is the identity mapping on $\mathbb{R}^N$. For any $\mathbf{y}$ we have $\mathbf{My} = \mathbf{Iy} - \mathbf{Py} = \mathbf{y} - \mathbf{Py}$ as required. In regression analysis $\mathbf{M}$ shows up as a matrix called the "annihilator." This is a pretty cool name, but it's also not a great description of its function. In what follows we will refer to $\mathbf{M}$ as the **residual projection**.

Example 2.2.3 Recall example 2.2.1, where we found that the projection of $\mathbf{y} \in \mathbb{R}^N$ onto span$\{\mathbf{1}\}$ is $\bar{y}\mathbf{1}$. The residual projection is $\mathbf{M}_c\mathbf{y} := \mathbf{y} - \bar{y}\mathbf{1}$. In econometric applications, we'll view this as a vector of errors obtained when the elements of a vector are predicted by its sample mean. The subscript reminds us that $\mathbf{M}_c$ centers vectors around their mean.

Fact 2.2.8 Let S be a linear subspace of $\mathbb{R}^N$, let $\mathbf{P} = \operatorname{proj} S$, and let $\mathbf{M}$ be the residual projection as defined in (2.11). The following statements are true:

(i) $\mathbf{M} = \operatorname{proj} S^{\perp}$.

(ii) $\mathbf{y} = \mathbf{Py} + \mathbf{My}$ for any $\mathbf{y} \in \mathbb{R}^N$.

(iii) $\mathbf{Py} \perp \mathbf{My}$ for any $\mathbf{y} \in \mathbb{R}^N$.

(iv) $\mathbf{My} = \mathbf{0}$ if and only if $\mathbf{y} \in S$.

(v) $\mathbf{P} \circ \mathbf{M} = \mathbf{M} \circ \mathbf{P} = \mathbf{0}$.

Part (v) means that $\mathbf{PMy} = \mathbf{MPy} = \mathbf{0}$ for all $\mathbf{y} \in \mathbb{R}^N$. Figure 2.17 illustrates the action of $\mathbf{M}$. The results in fact 2.2.8 can be seen in the figure.

If S_1 and S_2 are two subspaces of $\mathbb{R}^N$ with $S_1 \subset S_2$, then $S_2^{\perp} \subset S_1^{\perp}$. This means that the result in fact 2.2.7 is reversed for $\mathbf{M}$.

Fact 2.2.9 Let S_1 and S_2 be two subspaces of $\mathbb{R}^N$ and let $\mathbf{y} \in \mathbb{R}^N$. Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be the projections onto $S_1^\perp$ and $S_2^\perp$ respectively. If $S_1 \subset S_2$, then

$$\mathbf{M}_1\mathbf{M}_2\mathbf{y} = \mathbf{M}_2\mathbf{M}_1\mathbf{y} = \mathbf{M}_2\mathbf{y}$$

As an application of the ideas above, let's now discuss a procedure called Gram–Schmidt orthogonalization, which provides a fundamental link between the two major concepts discussed in this chapter: linear independence and orthogonality. It can be considered as a partial converse to fact 2.2.2 on page 27.

Theorem 2.2.3 *For each linearly independent set* $\{\mathbf{b}_1, \ldots, \mathbf{b}_K\} \subset \mathbb{R}^N$, *there exists an orthonormal set* $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ *with*

$$\operatorname{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_k\} = \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\} \quad \textit{for} \;\; k = 1, \ldots, K$$

The proof of theorem 2.2.3 provides an important algorithm for generating the orthonormal set $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$. The first step is to construct orthogonal sets $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ with span identical to $\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$ for each k. The construction of $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ uses the so called **Gram–Schmidt orthogonalization** procedure. First, for each $k = 1, \ldots, K$, let

(i) $B_k := \operatorname{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$,

(ii) $\mathbf{P}_k := \operatorname{proj} B_k$ and $\mathbf{M}_k := \operatorname{proj} B_k^\perp$,

(iii) $\mathbf{v}_k := \mathbf{M}_{k-1}\mathbf{b}_k$ where $\mathbf{M}_0$ is the identity mapping, and

(iv) $V_k := \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$.

In step (iii) we map each successive element $\mathbf{b}_k$ into a subspace orthogonal to the subspace generated by $\mathbf{b}_1, \ldots, \mathbf{b}_{k-1}$. In the exercises you are asked to show that $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ is an orthogonal set and $V_k = B_k$ for all k (ex. 2.4.34). To complete the argument, we introduce the vectors $\mathbf{u}_k$ by $\mathbf{u}_k := \mathbf{v}_k / \|\mathbf{v}_k\|$ and confirm that this set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is orthonormal with span equal to V_k. These results are also (solved) exercises.

2.3 Further Reading

Good texts on vector spaces include Marcus and Minc (1988) and Jänich (1994).

2.4 Exercises

Ex. 2.4.1 Show that inner products of linear combinations satisfy the following rule:

$$\left\langle \sum_{k=1}^{K} \alpha_k \mathbf{x}_k, \sum_{j=1}^{J} \beta_j \mathbf{y}_j \right\rangle = \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_k \beta_j \langle \mathbf{x}_k, \mathbf{y}_j \rangle$$

Ex. 2.4.2 Show that the vectors $(1, 1)$ and $(-1, 2)$ are linearly independent.

Ex. 2.4.3 Use fact 2.1.2 on page 13 to show that if $\mathbf{y} \in \mathbb{R}^N$ is such that $\langle \mathbf{y}, \mathbf{x} \rangle = 0$ for every $\mathbf{x} \in \mathbb{R}^N$, then $\mathbf{y} = \mathbf{0}$.

Ex. 2.4.4 Fix nonzero $\mathbf{x} \in \mathbb{R}^N$. Consider the optimization problem

$$\max_{\mathbf{y}} \langle \mathbf{x}, \mathbf{y} \rangle \quad \text{subject to} \quad \mathbf{y} \in \mathbb{R}^N \text{ and } \|\mathbf{y}\| = 1$$

Show that the maximizer is $\hat{\mathbf{x}} := (1/\|\mathbf{x}\|)\mathbf{x}$.[3]

Ex. 2.4.5 Is $\mathbb{R}^2$ a linear subspace of $\mathbb{R}^3$? Why or why not?

Ex. 2.4.6 Show that if $T : \mathbb{R}^K \to \mathbb{R}^N$ is a linear function, then $\mathbf{0} \in \ker T$.

Ex. 2.4.7 Let $\{\mathbf{x}_1, \mathbf{x}_2\}$ be a linearly independent set in $\mathbb{R}^2$ and let γ be a nonzero scalar. Is it true that $\{\gamma\mathbf{x}_1, \gamma\mathbf{x}_2\}$ is also linearly independent?

Ex. 2.4.8 Is it true that

$$\mathbf{z} := \begin{pmatrix} -3.9 \\ 12.4 \\ -6.8 \end{pmatrix} \in \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} \quad \text{when} \quad \mathbf{x}_1 = \begin{pmatrix} -4 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{x}_2 = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \ \mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}$$

Why or why not?

Ex. 2.4.9 Show that if S and S' are two linear subspaces of $\mathbb{R}^N$, then $S \cap S'$ is also a linear subspace of $\mathbb{R}^N$.

Ex. 2.4.10 Prove fact 2.1.5 on page 20.

Ex. 2.4.11 Let $Q := \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 = x_1 + x_3\}$. Is Q a linear subspace of $\mathbb{R}^3$?

Ex. 2.4.12 Let $Q := \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 = 1\}$. Is Q a linear subspace of $\mathbb{R}^3$?

3. Hint: There's no need to go taking derivatives and setting them equal to zero. An easier proof exists. If you're stuck, consider the Cauchy–Schwarz inequality.

Ex. 2.4.13 Show that if $T \colon \mathbb{R}^N \to \mathbb{R}^N$ is a linear function and λ is any scalar, then $E := \{\mathbf{x} \in \mathbb{R}^N : T\mathbf{x} = \lambda \mathbf{x}\}$ is a linear subspace of $\mathbb{R}^N$.

Ex. 2.4.14 Show that if $B \subset S$ with $\operatorname{span} B = S$, then $\mathbf{x} \perp S$ if and only if $\mathbf{x} \perp \mathbf{b}$ for all $\mathbf{b} \in B$.

Ex. 2.4.15 Prove the equivalences in theorem 2.1.1 on page 18.

Ex. 2.4.16 Prove fact 2.1.4 on page 19.

Ex. 2.4.17 Show that if S is a linear subspace of $\mathbb{R}^N$ then every basis of S has the same number of elements.

Ex. 2.4.18 Prove fact 2.1.6 on page 23.

Ex. 2.4.19 Prove theorem 2.1.6 on page 23.

Ex. 2.4.20 Show that if $T \colon \mathbb{R}^N \to \mathbb{R}^N$ is nonsingular (i.e., a linear bijection), then T^{-1} is also linear.

Ex. 2.4.21 Prove theorem 2.1.7 on page 25.

Ex. 2.4.22 Prove theorem 2.1.8 on page 26.

Ex. 2.4.23 Find two unit vectors (i.e., vectors with norm equal to one) that are orthogonal to $(1, -2)$.

Ex. 2.4.24 Prove the Pythagorean law (fact 2.2.1 on page 27). See ex. 2.4.1 if you need a hint.

Ex. 2.4.25 Prove fact 2.2.2 on page 27.

Ex. 2.4.26 Prove fact 2.2.8 using theorems 2.2.1 and 2.2.2.

Ex. 2.4.27 Prove fact 2.2.5: If $S \subset \mathbb{R}^N$, then $S \cap S^{\perp} = \{\mathbf{0}\}$.

Ex. 2.4.28 Prove fact 2.2.7 on page 32.

Ex. 2.4.29 Let $\mathbf{P}$ be the orthogonal projection described in theorem 2.2.2 (page 31). Confirm that $\mathbf{P}$ is a linear function from $\mathbb{R}^N$ to $\mathbb{R}^N$.

Ex. 2.4.30 Let $S := \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 = 0\}$ and let $\mathbf{y} := \mathbf{1} := (1, 1, 1)$. Using the orthogonal projection theorem, find the closest point in S to $\mathbf{y}$.

Ex. 2.4.31 Let S be any linear subspace of $\mathbb{R}^N$ and let $\mathbf{P} = \operatorname{proj} S$ (see theorem 2.2.2 on page 31). Is $\mathbf{P}$ one-to-one as a function on $\mathbb{R}^N$?

Ex. 2.4.32 Prove the reverse triangle inequality. That is, given two vectors $\mathbf{x}$ and $\mathbf{y}$, show that $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leqslant \|\mathbf{x} - \mathbf{y}\|$.[4]

Ex. 2.4.33 Show that $\mathbf{P}\mathbf{y} = \mathbf{0}$ implies $\mathbf{y} \in S^{\perp}$.

In the next three exercises, the notation is as given in theorem 2.2.3 and the discussion immediately afterwards.

Ex. 2.4.34 Show that $V_k = B_k$ for all k.

Ex. 2.4.35 Show that $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ is an orthogonal set.

Ex. 2.4.36 Show that $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is an orthonormal set with span equal to V_k for all k.

2.4.1 Solutions to Selected Exercises

Solution to Ex. 2.4.4. Fix nonzero $\mathbf{x} \in \mathbb{R}^N$. Let $\hat{\mathbf{x}} := \mathbf{x}/\|\mathbf{x}\|$. Comparing this point with any other $\mathbf{y} \in \mathbb{R}^N$ satisfying $\|\mathbf{y}\| = 1$, the Cauchy–Schwarz inequality yields

$$\langle \mathbf{y}, \mathbf{x} \rangle \leqslant | \langle \mathbf{y}, \mathbf{x} \rangle | \leqslant \|\mathbf{y}\| \|\mathbf{x}\| = \|\mathbf{x}\| = \frac{\langle \mathbf{x}, \mathbf{x} \rangle}{\|\mathbf{x}\|} = \langle \hat{\mathbf{x}}, \mathbf{x} \rangle$$

Hence $\hat{\mathbf{x}}$ is the maximizer, as claimed. □

Solution to Ex. 2.4.5. This is a bit of a trick question. To solve it, you need to look carefully at the definitions (as always). A linear subspace of $\mathbb{R}^3$ is a subset of $\mathbb{R}^3$ with certain properties. $\mathbb{R}^3$ is a collection of 3-tuples (x_1, x_2, x_3) where each x_i is a real number. Elements of $\mathbb{R}^2$ are 2-tuples (pairs), and hence not elements of $\mathbb{R}^3$. Therefore $\mathbb{R}^2$ is not a subset of $\mathbb{R}^3$, and, in particular, not a linear subspace of $\mathbb{R}^3$. □

Solution to Ex. 2.4.6. Let T be as in the question. We need to show that $T\mathbf{0} = \mathbf{0}$. Here's one proof. We know from the definition of scalar multiplication that $0\mathbf{x} = \mathbf{0}$ for any vector $\mathbf{x}$. Let $\mathbf{x}$ and $\mathbf{y}$ be any vectors in $\mathbb{R}^K$ and apply the definition of linearity to obtain

$$T\mathbf{0} = T(0\mathbf{x} + 0\mathbf{y}) = 0T\mathbf{x} + 0T\mathbf{y} = \mathbf{0} + \mathbf{0} = \mathbf{0}$$ □

Solution to Ex. 2.4.7. The answer is yes. Suppose, to the contrary, that $\{\gamma\mathbf{x}_1, \gamma\mathbf{x}_2\}$ is linearly dependent. Then one element can be written as a linear combination of the others. In our setting with only two vectors, this translates to $\gamma\mathbf{x}_1 = \alpha\gamma\mathbf{x}_2$ for some α. Since $\gamma \neq 0$, we can multiply each side by $1/\gamma$ to get $\mathbf{x}_1 = \alpha\mathbf{x}_2$. This contradicts linear independence of $\{\mathbf{x}_1, \mathbf{x}_2\}$. □

4. Hint: Use the triangle inequality.

Solution to Ex. 2.4.8. There is an easy way to do this: We know that any linearly independent set of 3 vectors in $\mathbb{R}^3$ will span $\mathbb{R}^3$. Since $\mathbf{z} \in \mathbb{R}^3$, this will include $\mathbf{z}$. So all we need to do is show that the 3 vectors are linearly independent. To this end, take any scalars $\alpha_1, \alpha_2, \alpha_3$ with

$$\alpha_1 \begin{pmatrix} -4 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} = \mathbf{0} := \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Written as three equations, this says that $-4\alpha_1 = 0$, $2\alpha_2 = 0$, and $-1\alpha_3 = 0$. Hence $\alpha_1 = \alpha_2 = \alpha_3 = 0$, and therefore the set is linearly independent. □

Solution to Ex. 2.4.9. Let S and S' be two linear subspaces of $\mathbb{R}^N$. Fix $\mathbf{x}, \mathbf{y} \in S \cap S'$ and $\alpha, \beta \in \mathbb{R}$. We claim that $\mathbf{z} := \alpha\mathbf{x} + \beta\mathbf{y} \in S \cap S'$. To see this, note that since $\mathbf{x}, \mathbf{y} \in S$ and S is a linear subspace, we have $\mathbf{z} \in S$; and since $\mathbf{x}, \mathbf{y} \in S'$ and S' is a linear subspace, we have $\mathbf{z} \in S'$. It follows that $\mathbf{z} \in S \cap S'$, as was to be shown. □

Solution to Ex. 2.4.11. If $\mathbf{a} := (1, -1, 1)$, then Q is all $\mathbf{x}$ with $\langle \mathbf{a}, \mathbf{x} \rangle = 0$. This set is a linear subspace of $\mathbb{R}^3$, as shown in example 2.1.10. □

Solution to Ex. 2.4.15. We are asked to verify the equivalences in theorem 2.1.1 on page 18 for the set $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$. We will prove the cycle (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (i).

((i) $\implies$ (ii)) We aim to show that if (i) holds and X_0 is a proper subset of X, then $\operatorname{span} X_0$ is a proper subset of $\operatorname{span} X$. To simplify notation let's take $X_0 := \{\mathbf{x}_2, \ldots, \mathbf{x}_K\}$. Suppose, to the contrary, that $\operatorname{span} X_0 = \operatorname{span} X$. Since $\mathbf{x}_1 \in \operatorname{span} X$, we must then have $\mathbf{x}_1 \in \operatorname{span} X_0$, from which we deduce the existence of scalars $\alpha_2, \ldots, \alpha_K$ such that $\mathbf{0} = -\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_K\mathbf{x}_K$. Since $-1 \neq 0$, this contradicts part (i).

((ii) $\implies$ (iii)) The claim is that when (ii) holds, no vector in X can be written as a linear combination of the others. Suppose, to the contrary, that $\mathbf{x}_1 = \alpha_2\mathbf{x}_2 + \cdots + \alpha_K\mathbf{x}_K$, say. Let $\mathbf{y} \in \operatorname{span} X$, so that $\mathbf{y} = \beta_1\mathbf{x}_1 + \cdots + \beta_K\mathbf{x}_K$. If we use the preceding equality to substitute out $\mathbf{x}_1$, we get $\mathbf{y}$ as a linear combination of $\{\mathbf{x}_2, \ldots, \mathbf{x}_K\}$ alone. In other words, any element of $\operatorname{span} X$ is in the span of the proper subset $\{\mathbf{x}_2, \ldots, \mathbf{x}_K\}$. Contradiction.

((iii) $\implies$ (i)) The final claim is that $\alpha_1 = \cdots = \alpha_K = 0$ whenever $\alpha_1\mathbf{x}_1 + \cdots + \alpha_K\mathbf{x}_K = \mathbf{0}$. Suppose, to the contrary, that there exist scalars with $\alpha_1\mathbf{x}_1 + \cdots + \alpha_K\mathbf{x}_K = \mathbf{0}$ and yet $\alpha_k \neq 0$ for at least one k. It follows immediately that $\mathbf{x}_k = (1/\alpha_k)\sum_{j \neq k} \alpha_j\mathbf{x}_j$. This contradicts (iii). □

Solution to Ex. 2.4.16. The aim is to prove fact 2.1.4 on page 19. Regarding the part (i), let's take X as linearly independent and show that the subset $X_0 := \{\mathbf{x}_1, \ldots, \mathbf{x}_{K-1}\}$ is linearly independent. (The argument for more general subsets is similar.) Suppose, to the contrary, that X_0 is linearly dependent. Then, by the definition, we can take $\alpha_1, \ldots, \alpha_{K-1}$ not all zero with $\sum_{k=1}^{K-1} \alpha_k \mathbf{x}_k = \mathbf{0}$. Letting $\alpha_K = 0$, we can write this as $\sum_{k=1}^{K} \alpha_k \mathbf{x}_k = \mathbf{0}$. Since not all coefficients are zero, we have contradicted linear independence of X.

Regarding (ii), let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ be linearly independent and suppose that $\mathbf{x}_j = \mathbf{0}$. Then by setting $\alpha_k = 0$ for $k \neq j$ and $\alpha_j = 1$, we can form scalars not all equal to zero with $\sum_{k=1}^{K} \alpha_k \mathbf{x}_k = \mathbf{0}$.

Regarding (iii), let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$ be linearly independent and let $\mathbf{x}_{K+1}$ be any point in $\mathbb{R}^N$ such that $\mathbf{x}_{K+1} \notin \operatorname{span} X$. The claim is that $X \cup \{\mathbf{x}_{K+1}\}$ is linearly independent. Suppose, to the contrary, that there exist $\alpha_1, \ldots, \alpha_K, \alpha_{K+1}$ not all zero such that $\sum_{k=1}^{K+1} \alpha_k \mathbf{x}_k = \mathbf{0}$. There are two possibilities for α_{K+1}, both of which lead to a contradiction: First, if $\alpha_{K+1} = 0$, then, since $\alpha_1, \ldots, \alpha_K, \alpha_{K+1}$ are not all zero, at least one of $\alpha_1, \ldots, \alpha_K$ are nonzero, and moreover $\sum_{k=1}^{K} \alpha_k \mathbf{x}_k = \sum_{k=1}^{K+1} \alpha_k \mathbf{x}_k = \mathbf{0}$. This contradicts our assumption of independence on X. Second, if $\alpha_{K+1} \neq 0$, then from $\sum_{k=1}^{K+1} \alpha_k \mathbf{x}_k = \mathbf{0}$ we can express $\mathbf{x}_{K+1}$ as a linear combination of elements of X. This contradicts the hypothesis that $\mathbf{x}_{K+1} \notin \operatorname{span} X$. □

Solution to Ex. 2.4.17. Let B_1 and B_2 be two bases of S, with K_1 and K_2 elements respectively. By definition, B_2 is a linearly independent subset of S. Moreover, S is spanned by the set B_1, which has K_1 elements. Applying theorem 2.1.3, we see that B_2 has at most K_1 elements. That is, $K_2 \leqslant K_1$. Reversing the roles of B_1 and B_2 gives $K_1 \leqslant K_2$. □

Solution to Ex. 2.4.18. The aim is to prove fact 2.1.6 on page 23. Suppose that S and S' are K-dimensional linear subspaces of $\mathbb{R}^N$ with $S \subset S'$. We claim that $S = S'$. To see this, observe that by the definition of dimension, S is equal to $\operatorname{span} B$ where B is a set of K linearly independent basis vectors $\{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$. If $S \neq S'$, then there exists a vector $\mathbf{x} \in S'$ such that $\mathbf{x} \notin \operatorname{span} B$. In view of theorem 2.1.1 on page 18, the set $\{\mathbf{x}, \mathbf{b}_1, \ldots, \mathbf{b}_K\}$ is linearly independent. Moreover, since $\mathbf{x} \in S'$ and since $B \subset S \subset S'$, we now have $K + 1$ linearly independent vectors inside S'. At the same time, being K-dimensional, we know that S' is spanned by K vectors. This contradicts theorem 2.1.3 on page 20.

Regarding part (ii), suppose that S is an M-dimensional linear subspace of $\mathbb{R}^N$ where $M < N$ and yet $S = \mathbb{R}^N$. Then we have a space S spanned by $M < N$ vectors that also contains the N linearly independent canonical basis vectors. We are led to another contradiction of theorem 2.1.3. Hence $S = \mathbb{R}^N$ cannot hold. □

Solution to Ex. 2.4.19. Regarding part (i), let B be a basis of $\operatorname{span} X$. By definition, B is a linearly independent subset of $\operatorname{span} X$. Since $\operatorname{span} X$ is spanned by K vectors, theorem 2.1.3 implies that B has no more than K elements. Hence, $\dim \operatorname{span} X \leqslant K$.

Regarding part (ii), suppose first that X is linearly independent. Then X is a basis for $\operatorname{span} X$. Since X has K elements, we conclude that $\dim \operatorname{span} X = K$. Conversely, if $\dim \operatorname{span} X = K$, then X must be linearly independent. If X were not linearly independent, then there would exist a proper subset X_0 of X such that $\operatorname{span} X_0 = \operatorname{span} X$. By part (i) of this theorem, we then have $\dim \operatorname{span} X_0 \leqslant \# X_0 \leqslant K - 1$. Therefore $\dim \operatorname{span} X \leqslant K - 1$, a contradiction. □

Solution to Ex. 2.4.20. Let $T \colon \mathbb{R}^N \to \mathbb{R}^N$ be nonsingular, and let T^{-1} be its inverse. To see that T^{-1} is linear, we need to show that for any pair $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^N$ (which is the domain of T^{-1}) and any scalars α and β, the following equality holds:

$$T^{-1}(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha T^{-1} \mathbf{x} + \beta T^{-1} \mathbf{y} \tag{2.12}$$

In the proof we will exploit the fact that T is by assumption a linear bijection.

Pick any vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and any two scalars α, β. Since T is a bijection, we know that $\mathbf{x}$ and $\mathbf{y}$ have unique preimages under T. In particular, there exist unique vectors $\mathbf{u}$ and $\mathbf{v}$ such that $T\mathbf{u} = \mathbf{x}$ and $T\mathbf{v} = \mathbf{y}$. Using these definitions, linearity of T and the fact that T^{-1} is the inverse of T, we have

$$T^{-1}(\alpha \mathbf{x} + \beta \mathbf{y}) = T^{-1}(\alpha T\mathbf{u} + \beta T\mathbf{v}) = T^{-1}(T(\alpha \mathbf{u} + \beta \mathbf{v})) = \alpha \mathbf{u} + \beta \mathbf{v} = \alpha T^{-1}\mathbf{x} + \beta T^{-1}\mathbf{y}$$

This chain of equalities confirms (2.12). □

Solution to Ex. 2.4.21. A collection of equivalent statements such as this is usually proved via a cycle of implications, with the form (i) $\implies$ (ii) $\implies \cdots \implies$ (vi) $\implies$ (i). However, in this case the logic is clearer if we directly show that all statements are equivalent to linear independence of V.

First observe equivalence of the onto property and linear independence of V via

$$T \text{ onto} \iff \operatorname{rng} T = \mathbb{R}^N \iff \operatorname{span} V = \mathbb{R}^N$$

by fact 2.1.7. The last statement is equivalent to linear independence of V by theorem 2.1.4 on page 22.

Next let's show that $\operatorname{null} T = \{\mathbf{0}\}$ implies linear independence of V. To this end, suppose that $\operatorname{null} T = \{\mathbf{0}\}$ and let $\alpha_1, \ldots, \alpha_N$ be such that $\sum_{n=1}^N \alpha_n T\mathbf{e}_n = \mathbf{0}$. By linearity of T, we then have $T(\sum_{n=1}^N \alpha_n \mathbf{e}_n) = \mathbf{0}$. Since $\operatorname{null} T = \{\mathbf{0}\}$, this means that $\sum_{n=1}^N \alpha_n \mathbf{e}_n = \mathbf{0}$, which in view of independence of $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$, implies $\alpha_1 = \cdots = \alpha_N = 0$. This establishes that V is linearly independent.

Now let's check that linear independence of V implies $\operatorname{null} T = \{\mathbf{0}\}$. To this end, let $\mathbf{x}$ be a vector in $\mathbb{R}^N$ such that $T\mathbf{x} = \mathbf{0}$. We can represent $\mathbf{x}$ in the form $\sum_{n=1}^N \alpha_n \mathbf{e}_n$ for suitable scalars $\{\alpha_n\}$. From linearity and $T\mathbf{x} = \mathbf{0}$, we get $\sum_{n=1}^N \alpha_n T\mathbf{e}_n = \mathbf{0}$. By linear independence of V, this implies that each $\alpha_n = 0$, whence $\mathbf{x} = \mathbf{0}$. Thus null $T = \{\mathbf{0}\}$ as claimed.

From fact 2.1.8 we have null $T = \{\mathbf{0}\}$ iff T is one-to-one, so we can now state the following equivalences

$$T \text{ onto} \iff V \text{ linearly independent} \iff T \text{ one-to-one} \tag{2.13}$$

If T is a bijection, then T is onto and hence V is linearly independent by (2.13). Conversely, if V is linearly independent then T is both onto and one-to-one by (2.13). Hence T is a bijection.

Finally, equivalence of linear independence of V and the statement that V forms a basis of $\mathbb{R}^N$ is immediate from the definition of bases and theorem 2.1.4 on page 22. □

Solution to Ex. 2.4.22. Regarding part (i), let $K < N$ and let $T \colon \mathbb{R}^K \to \mathbb{R}^N$ be linear. T cannot be onto because, if T were onto, then we would have rng $T = \mathbb{R}^N$, in which case the vectors in $V = \{T\mathbf{e}_1, \ldots, T\mathbf{e}_K\}$ in fact 2.1.7 would span $\mathbb{R}^N$, despite having only $K < N$ elements. This is impossible. (Why?)

Regarding part (ii), let $T \colon \mathbb{R}^K \to \mathbb{R}^N$ be linear and let $K > N$. Seeking a contradiction, suppose in addition that T is one-to-one. Let $\{\alpha_k\}_{k=1}^K$ be such that $\sum_{k=1}^K \alpha_k T\mathbf{e}_k = \mathbf{0}$. By linearity, $T(\sum_{k=1}^K \alpha_k \mathbf{e}_k) = \mathbf{0}$, and since T is one-to-one and $T\mathbf{0} = \mathbf{0}$, this in turn implies $\sum_{k=1}^K \alpha_k \mathbf{e}_k = \mathbf{0}$. Since the canonical basis vectors are linearly independent, it must be that $\alpha_1 = \cdots = \alpha_K = 0$. From this we conclude that $\{T\mathbf{e}_1, \ldots, T\mathbf{e}_K\}$ is linearly independent. Thus $\mathbb{R}^N$ contains K linearly independent vectors, despite the fact that $N < K$. This is impossible by theorem 2.1.3 on page 20. □

Solution to Ex. 2.4.25. Let $O = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \subset \mathbb{R}^N$ be an orthogonal set that does not contain $\mathbf{0}$. Let $\alpha_1, \ldots, \alpha_K$ be such that $\sum_{k=1}^K \alpha_k \mathbf{x}_k = \mathbf{0}$. We claim that $\alpha_j = 0$ for any j. To see that this is so, fix j and take the inner product of both sides of $\sum_{k=1}^K \alpha_k \mathbf{x}_k = \mathbf{0}$ with respect to $\mathbf{x}_j$ to obtain $\alpha_j \|\mathbf{x}_j\|^2 = 0$. Since $\mathbf{x}_j \neq \mathbf{0}$, we conclude that $\alpha_j = 0$. The proof is done. □

Solution to Ex. 2.4.27. Let $S \subset \mathbb{R}^N$. We aim to show that $S \cap S^\perp = \{\mathbf{0}\}$. Fix $\mathbf{a} \in S \cap S^\perp$. Since $\mathbf{a} \in S^\perp$, we know that $\langle \mathbf{a}, \mathbf{s} \rangle = 0$ for any $\mathbf{s} \in S$. Since $\mathbf{a} \in S$, we have in particular, $\langle \mathbf{a}, \mathbf{a} \rangle = \|\mathbf{a}\|^2 = 0$. As we saw in fact 2.1.2, the only such vector is $\mathbf{0}$. □

Solution to Ex. 2.4.29. Fix $\alpha, \beta \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. The claim is that

$$\mathbf{P}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{P}\mathbf{x} + \beta\mathbf{P}\mathbf{y}$$

To verify this equality, we need to show that the right-hand side is the orthogonal projection of $\alpha\mathbf{x} + \beta\mathbf{y}$ onto S. Going back to theorem 2.2.1, we need to show that (i)

$\alpha \mathbf{P}\mathbf{x} + \beta \mathbf{P}\mathbf{y} \in S$ and (ii) for any $\mathbf{z} \in S$, we have

$$\langle \alpha \mathbf{x} + \beta \mathbf{y} - (\alpha \mathbf{P}\mathbf{x} + \beta \mathbf{P}\mathbf{y}), \mathbf{z} \rangle = 0$$

Here (i) is immediate, because $\mathbf{P}\mathbf{x}$ and $\mathbf{P}\mathbf{y}$ are in S by definition, and, moreover S is a linear subspace. To see that (ii) holds, just note that

$$\langle \alpha \mathbf{x} + \beta \mathbf{y} - (\alpha \mathbf{P}\mathbf{x} + \beta \mathbf{P}\mathbf{y}), \mathbf{z} \rangle = \alpha \langle \mathbf{x} - \mathbf{P}\mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y} - \mathbf{P}\mathbf{y}, \mathbf{z} \rangle$$

By definition, the projections of $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to S, so we have $\langle \mathbf{x} - \mathbf{P}\mathbf{x}, \mathbf{z} \rangle = \langle \mathbf{y} - \mathbf{P}\mathbf{y}, \mathbf{z} \rangle = 0$. We are done. □

Solution to Ex. 2.4.30. Let $\mathbf{x} = (x_1, x_2, x_3)$ be the closest point in S to $\mathbf{y}$. Note that $\mathbf{e}_1 \in S$ and $\mathbf{e}_2 \in S$. By the orthogonal projection theorem, we have (i) $\mathbf{x} \in S$ and (ii) $\mathbf{y} - \mathbf{x} \perp S$. From (i) we have $x_3 = 0$. From (ii) we have

$$\langle \mathbf{y} - \mathbf{x}, \mathbf{e}_1 \rangle = 0 \quad \text{and} \quad \langle \mathbf{y} - \mathbf{x}, \mathbf{e}_2 \rangle = 0$$

These equations can be expressed more simply as $1 - x_1 = 0$ and $1 - x_2 = 0$. We conclude that $\mathbf{x} = (1, 1, 0)$. □

Solution to Ex. 2.4.31. If $S = \mathbb{R}^N$, then $\mathbf{P}$ is the identity mapping (why?), which is one-to-one. If $S \neq \mathbb{R}^N$, then take any $\mathbf{x} \notin S$. By definition, $\mathbf{y} := \mathbf{P}\mathbf{x}$ is in S, and hence $\mathbf{y}$ and $\mathbf{x}$ are distinct. But $\mathbf{P}$ maps elements of S to themselves, so $\mathbf{P}\mathbf{y} = \mathbf{y} = \mathbf{P}\mathbf{x}$. Hence $\mathbf{P}$ is not one-to-one. □

Solution to Ex. 2.4.32. From the triangle inequality we have

$$\|\mathbf{x}\| = \|\mathbf{x} - \mathbf{y} + \mathbf{y}\| \leqslant \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y}\|$$

It follows that $\|\mathbf{x}\| - \|\mathbf{y}\| \leqslant \|\mathbf{x} - \mathbf{y}\|$. A similar argument reversing the roles of $\mathbf{x}$ and $\mathbf{y}$ gives $\|\mathbf{y}\| - \|\mathbf{x}\| \leqslant \|\mathbf{x} - \mathbf{y}\|$. Combining the last two inequalities gives

$$-\|\mathbf{x} - \mathbf{y}\| \leqslant \|\mathbf{x}\| - \|\mathbf{y}\| \leqslant \|\mathbf{x} - \mathbf{y}\|$$

This is equivalent to $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leqslant \|\mathbf{x} - \mathbf{y}\|$. □

Solution to Ex. 2.4.33. If $\mathbf{P}\mathbf{y} = \mathbf{0}$, then $\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{y}$. Hence $\mathbf{M}$ does not shift $\mathbf{y}$. If an orthogonal projection onto a subspace doesn't shift a point, that's because the point is already in that subspace (see theorem 2.2.2). In this case the subspace is $S^\perp$, and we conclude that $\mathbf{y} \in S^\perp$. □

Solution to Ex. 2.4.34. To see that $V_k = B_k$ for all k, fix k and consider the claim that $V_k \subset B_k$. By definition, $\mathbf{v}_k = \mathbf{b}_k - \mathbf{P}_{k-1}\mathbf{b}_k$, and the two terms on the right-hand side lie in B_k. Hence $\mathbf{v}_k \in B_k$. Since spans increase as we add more elements, it follows that $\mathbf{v}_j \in B_k$ for $j \leqslant k$. In other words, $\{\mathbf{v}_1, \dots, \mathbf{v}_k\} \subset B_k$. Since B_k is a linear subspace, we have $V_k \subset B_k$.

An induction argument shows that $B_k \subset V_k$ also holds. Clearly it holds for $k = 1$. Suppose that it also holds for $k - 1$. From the definition of $\mathbf{v}_k$, we have $\mathbf{b}_k = \mathbf{P}_{k-1}\mathbf{b}_k + \mathbf{v}_k$. The first term on the right-hand side lies in B_{k-1}, which, by our induction hypothesis, satisfies $B_{k-1} \subset V_{k-1} \subset V_k$. The second term on the right-hand side is $\mathbf{v}_k$, which obviously lies in V_k. Hence both terms are in V_k, and therefore $\mathbf{b}_k \in V_k$. Using arguments analogous to the end of the last paragraph leads us to $B_k \subset V_k$. □

Solution to Ex. 2.4.35. To show that $\{\mathbf{v}_1, \dots, \mathbf{v}_K\}$ is an orthogonal set, it suffices to check that $\mathbf{v}_k \perp \mathbf{v}_j$ whenever $j < k$. To see this, fix any pair $j < k$. By construction, $\mathbf{v}_k \in B_{k-1}^{\perp}$. But, as shown in the solution to ex. 2.4.34, $\mathbf{v}_j \in B_{k-1}$ must hold. Hence $\mathbf{v}_k \perp \mathbf{v}_j$. □

Solution to Ex. 2.4.36. Since $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is orthogonal, the family $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ will be orthonormal provided that the norm of each element is 1. This is true by construction, since $\mathbf{u}_k := \mathbf{v}_k / \|\mathbf{v}_k\|$. The only concern is that $\|\mathbf{v}_k\| = 0$ might hold for some k. But $\mathbf{v}_k = \mathbf{0}$ is impossible because, if it was to hold, then by (vii) of theorem 2.2.2 we would have $\mathbf{b}_k \in B_{k-1}$, contradicting linear independence of $\{\mathbf{b}_1, \dots, \mathbf{b}_K\}$.

The proof that $\operatorname{span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\} = V_k$ is straightforward and left for you. □