Chapter 3

Linear Algebra and Matrices

The previous chapter was relatively abstract but it sets us up nicely to study the kinds of practical problems from linear algebra that we encounter in statistics and econometrics. This chapter treats many of the core problems in linear algebra, frequently relating them back to our analysis of vectors and linear maps in chapter 2.

3.1 Matrices and Linear Equations

Matrices provide a convenient way to organize data and algebraic operations. Let's start with definitions.

3.1.1 Basic Definitions

A $N \times K$ **matrix** is a rectangular array $\mathbf{A}$ of real numbers with N rows and K columns, written in the following way:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{pmatrix}$$

The symbol a_{nk} stands for the element in the nth row of the kth column. Often these elements represent coefficients in a system of linear equations, such as

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1K}x_K &= b_1 \\ &\vdots \\ a_{N1}x_1 + a_{N2}x_2 + \cdots + a_{NK}x_K &= b_N \end{aligned} \tag{3.1}$$

For obvious reasons, $\mathbf{A}$ is also called a vector if either $N = 1$ or $K = 1$. In the former case, $\mathbf{A}$ is called a **row vector**, while in the latter case it is called a **column vector**. When convenient, we will use the notation $\text{row}_n \mathbf{A}$ to refer to the nth row of $\mathbf{A}$, and $\text{col}_k \mathbf{A}$ to refer to its kth column.

We extend the notation $\mathbf{0}$ and $\mathbf{1}$ from vectors to matrices, in the sense that these symbols will also represent matrices with all elements equal to zero and one respectively. Dimensions will be stated explicitly or clear from context.

If $\mathbf{A}$ is $N \times K$ and $N = K$, then $\mathbf{A}$ is called **square**. For an $N \times N$ matrix $\mathbf{A}$, the N elements of the form a_{nn} for $n = 1, \ldots, N$ are called the **principal diagonal**:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix}$$

The unique $N \times N$ matrix with ones on the principle diagonal and zeros elsewhere is called the **identity matrix**, and denoted by $\mathbf{I}$:

$$\mathbf{I} := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Note that $\text{col}_n \mathbf{I} = \mathbf{e}_n$, the nth canonical basis vector in $\mathbb{R}^N$.

Just as was the case for vectors, a number of algebraic operations are defined for matrices. The first two, scalar multiplication and addition, are immediate generalizations of the vector case: For $\gamma \in \mathbb{R}$, we let

$$\gamma \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{pmatrix} := \begin{pmatrix} \gamma a_{11} & \gamma a_{12} & \cdots & \gamma a_{1K} \\ \gamma a_{21} & \gamma a_{22} & \cdots & \gamma a_{2K} \\ \vdots & \vdots & & \vdots \\ \gamma a_{N1} & \gamma a_{N2} & \cdots & \gamma a_{NK} \end{pmatrix}$$

while

$$\begin{pmatrix} a_{11} & \cdots & a_{1K} \\ a_{21} & \cdots & a_{2K} \\ \vdots & \vdots & \vdots \\ a_{N1} & \cdots & a_{NK} \end{pmatrix} + \begin{pmatrix} b_{11} & \cdots & b_{1K} \\ b_{21} & \cdots & b_{2K} \\ \vdots & \vdots & \vdots \\ b_{N1} & \cdots & b_{NK} \end{pmatrix} := \begin{pmatrix} a_{11}+b_{11} & \cdots & a_{1K}+b_{1K} \\ a_{21}+b_{21} & \cdots & a_{2K}+b_{2K} \\ \vdots & \vdots & \vdots \\ a_{N1}+b_{N1} & \cdots & a_{NK}+b_{NK} \end{pmatrix}$$

Addition is only defined for matrices that have identical shape.

The **matrix product** $\mathbf{C} = \mathbf{AB}$ of matrices $\mathbf{A}$ and $\mathbf{B}$ is formed by taking as its i, jth element the inner product of the ith row of $\mathbf{A}$ and the jth column of $\mathbf{B}$. That is,

$$c_{ij} = \langle \operatorname{row}_i \mathbf{A}, \operatorname{col}_j \mathbf{B} \rangle = \sum_{k=1}^{K} a_{ik} b_{kj}$$

Here's the picture for $i = j = 1$:

$$\begin{pmatrix} a_{11} & \cdots & a_{1K} \\ a_{21} & \cdots & a_{2K} \\ \vdots & \vdots & \vdots \\ a_{N1} & \cdots & a_{NK} \end{pmatrix} \begin{pmatrix} b_{11} & \cdots & b_{1J} \\ b_{21} & \cdots & b_{2J} \\ \vdots & \vdots & \vdots \\ b_{K1} & \cdots & b_{KJ} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1J} \\ c_{21} & \cdots & c_{2J} \\ \vdots & \vdots & \vdots \\ c_{N1} & \cdots & c_{NJ} \end{pmatrix}$$

Since inner products are only defined for vectors of equal length, we need the length of the rows of $\mathbf{A}$ to be equal to the length of the columns of $\mathbf{B}$. In other words, if $\mathbf{A}$ is $N \times K$ and $\mathbf{B}$ is $J \times M$, then we require $K = J$. The resulting matrix $\mathbf{AB}$ is $N \times M$. Here's the rule to remember:

$$\mathbf{A} \text{ is } N \times K \text{ and } \mathbf{B} \text{ is } K \times M \implies \mathbf{AB} \text{ is } N \times M$$

Matrix multiplication is not commutative: $\mathbf{AB}$ and $\mathbf{BA}$ are not in general equal. In most other ways it behaves like ordinary multiplication:

Fact 3.1.1 For conformable matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and scalar α, we have

(i) $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$,

(ii) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$,

(iii) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$,

(iv) $\mathbf{A}\alpha\mathbf{B} = \alpha\mathbf{AB}$, and

(v) $\mathbf{AI} = \mathbf{A}$ and $\mathbf{IA} = \mathbf{A}$, where $\mathbf{I}$ is the identity matrix.

Here and below we use the word **conformable** to indicate dimensions are such that the operation in question makes sense. For example, two matrices are conformable for addition if they have the same number of rows and columns.

The **kth power** of a square matrix $\mathbf{A}$ is defined as

$$\mathbf{A}^k := \underbrace{\mathbf{A} \cdots \mathbf{A}}_{k \text{ terms}}$$

If $\mathbf{B}$ is such that $\mathbf{B}^2 = \mathbf{A}$, then $\mathbf{B}$ is called the **square root** of $\mathbf{A}$ and written $\sqrt{\mathbf{A}}$.

Before going further, let's state the fundamental connection between matrix multiplication and the more elementary notion of linear combinations of vectors, as introduced in §2.1.2: Given $N \times K$ matrix $\mathbf{A}$ and $K \times 1$ column vector $\mathbf{x}$, the product $\mathbf{Ax}$ is an $N \times 1$ column vector formed as a linear combination of the columns of $\mathbf{A}$, with scalars $x_1, \ldots, x_K$. In symbols,

$$\begin{aligned}
\mathbf{Ax} &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix} \\
&= x_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{N1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{N2} \end{pmatrix} + \cdots + x_K \begin{pmatrix} a_{1K} \\ a_{2K} \\ \vdots \\ a_{NK} \end{pmatrix} \\
&= \sum_{k=1}^{K} x_k \operatorname{col}_k \mathbf{A}
\end{aligned}$$

3.1.2 Matrices as Maps

One of the most useful ways to think about matrices is as maps from one vector space to another. In particular, an $N \times K$ matrix $\mathbf{A}$ can be thought of as a map sending a vector $\mathbf{x} \in \mathbb{R}^K$ into a new vector $\mathbf{y} = \mathbf{Ax}$ in $\mathbb{R}^N$. As the next theorem shows, these maps are always linear. In fact, they are the *only* linear functions. In other words, the set of linear functions from $\mathbb{R}^K$ to $\mathbb{R}^N$ and the set of $N \times K$ matrices are in one-to-one correspondence:

Theorem 3.1.1 *Let T be a function from $\mathbb{R}^K$ to $\mathbb{R}^N$. The following are equivalent:*

(i) T is linear.

(ii) There exists an $N \times K$ matrix $\mathbf{A}$ such that $T\mathbf{x} = \mathbf{Ax}$ for all $\mathbf{x} \in \mathbb{R}^K$.

Proof. ((i) $\implies$ (ii)) Let $T \colon \mathbb{R}^K \to \mathbb{R}^N$ be linear. We aim to construct a matrix $\mathbf{A}$ such that $T\mathbf{x} = \mathbf{Ax}$ for all $\mathbf{x} \in \mathbb{R}^K$. As usual, let $\mathbf{e}_k$ be the kth canonical basis vector in

$\mathbb{R}^K$. Define an $N \times K$ matrix $\mathbf{A}$ by $\operatorname{col}_k \mathbf{A} = T\mathbf{e}_k$. Pick any $\mathbf{x} \in \mathbb{R}^K$. We can also write $\mathbf{x} = \sum_{k=1}^K x_k \mathbf{e}_k$, where x_k is the kth element of $\mathbf{x}$. By linearity, we have

$$T\mathbf{x} = \sum_{k=1}^K x_k T\mathbf{e}_k = \sum_{k=1}^K x_k \operatorname{col}_k \mathbf{A}$$

This is just $\mathbf{A}\mathbf{x}$, as shown in §3.1.1.

((ii) $\implies$ (i)) Fix $N \times K$ matrix $\mathbf{A}$ and consider the function $T\colon \mathbb{R}^K \to \mathbb{R}^N$ defined by $T\mathbf{x} = \mathbf{A}\mathbf{x}$. Pick any $\mathbf{x}$, $\mathbf{y}$ in $\mathbb{R}^K$, and any scalars α and β. The rules of matrix arithmetic (see fact 3.1.1) tell us that

$$T(\alpha\mathbf{x} + \beta\mathbf{y}) := \mathbf{A}(\alpha\mathbf{x} + \beta\mathbf{y}) = \mathbf{A}\alpha\mathbf{x} + \mathbf{A}\beta\mathbf{y} = \alpha\mathbf{A}\mathbf{x} + \beta\mathbf{A}\mathbf{y} =: \alpha T\mathbf{x} + \beta T\mathbf{y}$$

Hence T is linear. □

When we consider the problem of solving a system of linear equations such as $\mathbf{A}\mathbf{x} = \mathbf{b}$, the first issue we need to concern ourselves with is existence. Can we find an $\mathbf{x}$ that satisfies this equation, for any given $\mathbf{b}$? A bit of thought will convince you that this is the same question as: Is the corresponding linear map $T\mathbf{x} = \mathbf{A}\mathbf{x}$ an onto function? (See §15.2 for the definition.) Equivalently, is $\operatorname{rng} T$ equal to all of $\mathbb{R}^N$?

The range of T is all vectors of the form $T\mathbf{x} = \mathbf{A}\mathbf{x}$ where $\mathbf{x}$ varies over $\mathbb{R}^K$. We just saw in §3.1.1 that, for any $\mathbf{x} \in \mathbb{R}^K$, we have $\mathbf{A}\mathbf{x} = \sum_{k=1}^K x_k \operatorname{col}_k \mathbf{A}$. It follows that $\operatorname{rng} T$ is equal to the **column space** of $\mathbf{A}$, which is by definition the span of the columns of $\mathbf{A}$. We represent it by the symbol

$$\operatorname{colspace} \mathbf{A} := \operatorname{span}\{\operatorname{col}_1 \mathbf{A}, \ldots, \operatorname{col}_K \mathbf{A}\} \tag{3.2}$$

To summarize the preceding discussion, we have

$$\operatorname{colspace} \mathbf{A} = \operatorname{rng} T = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^K\}$$

How large is the column space of a given matrix? To answer that question, we have to say what “large" means. In the context of linear algebra, size of subspaces is usually measured by dimension. The dimension of $\operatorname{colspace} \mathbf{A}$ is known as the **rank** of $\mathbf{A}$. That is,

$$\operatorname{rank} \mathbf{A} := \dim \operatorname{colspace} \mathbf{A}$$

$\mathbf{A}$ is said to have **full column rank** if $\operatorname{rank} \mathbf{A}$ is equal to K, the number of its columns. The reason we say “full” rank here is that, by definition, $\operatorname{colspace} \mathbf{A}$ is the span of K vectors. Hence, by part (i) of theorem 2.1.6 on page 23, we must have $\dim \operatorname{colspace} \mathbf{A} \leqslant K$. In other words, the rank of $\mathbf{A}$ is less than or equal to K. $\mathbf{A}$ is said to have full column rank when this maximum is achieved.

When is this maximum achieved? By part (ii) of theorem 2.1.6, this will be the case precisely when the columns of $\mathbf{A}$ are linearly independent. Let's state this as a fact:

Fact 3.1.2 Let $\mathbf{A}$ be an $N \times K$ matrix. The following statements are equivalent:

(i) $\mathbf{A}$ is of full column rank.

(ii) The columns of $\mathbf{A}$ are linearly independent.

(iii) $\mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{x} = \mathbf{0}$.

The last equivalence follows from theorem 2.1.1 on page 18.

3.1.3 Square Matrices and Invertibility

Perhaps the most common problem in linear algebra is solving systems of linear equations. A generic representation is $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{x}$ contains the unknowns and $\mathbf{A}$ and $\mathbf{b}$ are given. There are a variety of scenarios depending on the properties of $\mathbf{A}$. For now let's consider the case where $\mathbf{A}$ is $N \times N$, and seek conditions on $\mathbf{A}$ under which, for every $\mathbf{b} \in \mathbb{R}^N$, there exists exactly one $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{A}\mathbf{x} = \mathbf{b}$.

The best way to understand this problem is as follows. Let T be the linear map $T\mathbf{x} = \mathbf{A}\mathbf{x}$. The question we are asking here is: When does each point $\mathbf{b} \in \mathbb{R}^N$ have one and only one preimage under T? In other words, when is T a bijection?

To answer this question, we can refer back to the discussion in §2.1.7. We saw there that linear bijections are called nonsingular functions, and we discussed a number of equivalences for this property. The next fact replicates these equivalences in the language of matrices.

Fact 3.1.3 For $N \times N$ matrix $\mathbf{A}$, the following are equivalent:

(i) The columns of $\mathbf{A}$ are linearly independent.

(ii) The columns of $\mathbf{A}$ form a basis of $\mathbb{R}^N$.

(iii) $\operatorname{rank} \mathbf{A} = N$.

(iv) $\operatorname{colspace} \mathbf{A} = \mathbb{R}^N$.

(v) $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{y} \implies \mathbf{x} = \mathbf{y}$.

(vi) $\mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{x} = \mathbf{0}$.

(vii) For each $\mathbf{b} \in \mathbb{R}^N$, the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution.

(viii) For each $\mathbf{b} \in \mathbb{R}^N$, the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a unique solution.

These results can all be verified using theorem 2.1.7 on page 25 by checking the corresponding implications for $T\mathbf{x} = \mathbf{A}\mathbf{x}$. For example, it's easy to see that if $\mathbf{e}_n$ is the

nth canonical basis vector of $\mathbb{R}^N$, then

$$T\mathbf{e}_n = \mathbf{A}\mathbf{e}_n = \operatorname{col}_n \mathbf{A}$$

and hence the condition of linear independence of $\{T\mathbf{e}_1, \ldots, T\mathbf{e}_N\}$ in theorem 2.1.7 translates to linear independence of the columns of $\mathbf{A}$.

Following common usage, if any of the equivalent conditions in fact 3.1.3 are true we will call not just the map T but also the underlying matrix $\mathbf{A}$ **nonsingular**. If any one—and hence all—of these conditions fail, then $\mathbf{A}$ is called **singular**.

Any bijection has an inverse (see §15.2). In fact any nonsingular map T has a nonsingular inverse T^{-1} (see fact 2.1.9 on page 26). In the present setting, where T is generated by a matrix $\mathbf{A}$, the inverse T^{-1} is also associated with a matrix, called the inverse of $\mathbf{A}$. The next theorem gives details:

Theorem 3.1.2 *For nonsingular* $\mathbf{A}$ *the following statements are true:*

(i) There exists a square matrix $\mathbf{B}$ *such that* $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$*, where* $\mathbf{I}$ *is the identity matrix. The matrix* $\mathbf{B}$ *is called the* **inverse** *of* $\mathbf{A}$*, and written as* $\mathbf{A}^{-1}$*.*

(ii) For each $\mathbf{b} \in \mathbb{R}^N$*, the unique solution to* $\mathbf{Ax} = \mathbf{b}$ *is given by*

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{3.3}$$

For this reason, nonsingular matrices are also referred to as **invertible**. The proof of theorem 3.1.2 is a solved exercise (ex. 3.5.5).

Example 3.1.1 Consider the N good linear demand system

$$q_n = \sum_{k=1}^{N} a_{nk} p_k + b_n, \quad n = 1, \ldots, N$$

where q_n and p_n are quantity and price of the nth good respectively. We want to compute the inverse demand function, which gives prices in terms of quantities. To do so, we write our system in matrix form as $\mathbf{q} = \mathbf{Ap} + \mathbf{b}$. If the columns of $\mathbf{A}$ are linearly independent, then we can invert the system: a unique solution exists for each fixed $\mathbf{q}$ and $\mathbf{b}$. That solution is given by $\mathbf{p} = \mathbf{A}^{-1}(\mathbf{q} - \mathbf{b})$.

As stated in the next fact, to show that $\mathbf{A}$ is invertible and $\mathbf{B}$ is the inverse of $\mathbf{A}$, it suffices to show that $\mathbf{B}$ is either a **left inverse**, in the sense that $\mathbf{BA} = \mathbf{I}$, or a **right inverse**, in the sense that $\mathbf{AB} = \mathbf{I}$. A proof is given in §3.1.4.

Fact 3.1.4 Let $\mathbf{A}$ and $\mathbf{B}$ be $N \times N$ square matrices. If $\mathbf{B}$ is either a left or a right inverse of $\mathbf{A}$, then $\mathbf{A}$ is nonsingular and $\mathbf{B}$ is its inverse.

The next fact collects more useful results about inverse matrices.

Fact 3.1.5 If $\mathbf{A}$ and $\mathbf{B}$ are nonsingular and $\alpha \neq 0$, then

(i) $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$,

(ii) $(\alpha\mathbf{A})^{-1} = \alpha^{-1}\mathbf{A}^{-1}$, and

(iii) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

The relation $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ is just a special case of the analogous rule for inversion of bijections (see fact 15.2.1 on page 410). But you can also prove the equality directly by confirming that $\mathbf{B}^{-1}\mathbf{A}^{-1}$ is a right inverse (or left inverse) of $\mathbf{AB}$.

3.1.4 Determinants

To each square matrix $\mathbf{A}$, we can associate a unique number $\det \mathbf{A}$ called the determinant of $\mathbf{A}$. To define it, let $S(N)$ be the set of all bijections from $\{1, \ldots, N\}$ to itself. For $\pi \in S(N)$ we define the **signature** of π as

$$\operatorname{sgn}(\pi) := \prod_{m<n} \frac{\pi(m) - \pi(n)}{m - n}$$

The **determinant** of $N \times N$ matrix $\mathbf{A}$ is then given as

$$\det \mathbf{A} := \sum_{\pi \in S(N)} \operatorname{sgn}(\pi) \prod_{n=1}^{N} a_{\pi(n)n}$$

We won't concern ourselves with the details of this definition. For now it's enough to know the following facts:

Fact 3.1.6 If $\mathbf{I}$ is the $N \times N$ identity, $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ matrices and $\alpha \in \mathbb{R}$, then

(i) $\det \mathbf{I} = 1$,

(ii) $\mathbf{A}$ is nonsingular if and only if $\det \mathbf{A} \neq 0$,

(iii) $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$,

(iv) $\det(\alpha\mathbf{A}) = \alpha^N \det(\mathbf{A})$, and

(v) $\det(\mathbf{A}^{-1}) = (\det(\mathbf{A}))^{-1}$.

In the 2×2 case one can show that the determinant satisfies

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc \tag{3.4}$$

As an example of how these results can be useful, let's go back and prove fact 3.1.4. Fix square matrix $\mathbf{A}$ and suppose that a right inverse $\mathbf{B}$ exists, in the sense that $\mathbf{AB} = \mathbf{I}$. This equality implies that both $\mathbf{A}$ and $\mathbf{B}$ are nonsingular, since applying det to both sides of $\mathbf{AB} = \mathbf{I}$ and using the rules in fact 3.1.6 gives $\det(\mathbf{A})\det(\mathbf{B}) = 1$. It follows that both $\det \mathbf{A}$ and $\det \mathbf{B}$ are nonzero. Hence both matrices are nonsingular.

To show that $\mathbf{B}$ is the inverse of $\mathbf{A}$, we just need to check that, in addition to $\mathbf{AB} = \mathbf{I}$, we have $\mathbf{BA} = \mathbf{I}$. To obtain the latter equality from the former, premultiply the former by $\mathbf{B}$ to get $\mathbf{BAB} = \mathbf{B}$ and then postmultiply by $\mathbf{B}^{-1}$ to get $\mathbf{BA} = \mathbf{I}$. The proof for the left inverse case is similar.

3.2 Properties of Matrices

Let's look at some special kinds of matrices and their role in linear algebra.

3.2.1 Diagonal and Triangular Matrices

A square matrix is called

- **lower triangular** if each element strictly above the principal diagonal is zero,
- **upper triangular** if every element strictly below the principal diagonal is zero, and
- **triangular** if it is either upper or lower triangular.

For example, if we define

$$\mathbf{L} := \begin{pmatrix} 1 & 0 & 0 \\ 2 & 5 & 0 \\ 3 & 6 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{U} := \begin{pmatrix} 1 & 2 & 3 \\ 0 & 5 & 6 \\ 0 & 0 & 1 \end{pmatrix}$$

then $\mathbf{L}$ and $\mathbf{U}$ are lower and upper triangular respectively. The great advantage of triangular matrices is that the associated linear equations are simple to solve using either forward or backward substitution. For example, with the system

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 5 & 0 \\ 3 & 6 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ 2x_1 + 5x_2 \\ 3x_1 + 6x_2 + x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

the top equation involves only x_1, so we can solve for its value directly. Plugging this value into the second equation, we can solve for x_2 and so on.

Fact 3.2.1 If $\mathbf{A} = (a_{mn})$ is triangular, then $\det \mathbf{A} = \prod_{n=1}^{N} a_{nn}$.

An even better situation arises when our matrix is *both* upper and lower triangular. These are the **diagonal** matrices. In other words, a square matrix $\mathbf{A}$ is diagonal if all entries off the principal diagonal are zero. For example, the identity matrix is diagonal.

The following notation is often used to define diagonal matrices:

$$\operatorname{diag}(d_1, \ldots, d_N) := \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_N \end{pmatrix}$$

With diagonal matrices it is trivial to compute powers, roots, inverses and products:

Fact 3.2.2 Let $\mathbf{C} = \operatorname{diag}(c_1, \ldots, c_N)$ and $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_N)$. The following statements are true:

(i) $\mathbf{C} + \mathbf{D} = \operatorname{diag}(c_1 + d_1, \ldots, c_N + d_N)$.

(ii) $\mathbf{CD} = \operatorname{diag}(c_1 d_1, \ldots, c_N d_N)$.

(iii) $\mathbf{D}^k = \operatorname{diag}(d_1^k, \ldots, d_N^k)$ for any $k \in \mathbb{N}$.

(iv) If $d_n \geqslant 0$ for all n, then $\sqrt{\mathbf{D}}$ exists and equals $\operatorname{diag}(\sqrt{d_1}, \ldots, \sqrt{d_N})$.

(v) If $d_n \neq 0$ for all n, then $\mathbf{D}$ is nonsingular and $\mathbf{D}^{-1} = \operatorname{diag}(d_1^{-1}, \ldots, d_N^{-1})$.

You can check parts (i) and (ii) directly. The other claims follow from (i) and (ii).

3.2.2 Trace, Transpose, and Symmetry

The **trace** of an $N \times N$ matrix $\mathbf{A}$ is the sum of the elements on its principal diagonal:

$$\operatorname{trace} \mathbf{A} = \sum_{n=1}^{N} a_{nn}$$

Fact 3.2.3 If $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ matrices and α and β are two scalars, then

$$\operatorname{trace}(\alpha \mathbf{A} + \beta \mathbf{B}) = \alpha \operatorname{trace}(\mathbf{A}) + \beta \operatorname{trace}(\mathbf{B})$$

Moreover, if $\mathbf{A}$ is $N \times M$ and $\mathbf{B}$ is $M \times N$, then $\operatorname{trace}(\mathbf{AB}) = \operatorname{trace}(\mathbf{BA})$.

The **transpose** of $N \times K$ matrix $\mathbf{A}$ is a $K \times N$ matrix $\mathbf{A}^\top$ such that $\operatorname{col}_n(\mathbf{A}^\top) =$

$\text{row}_n \mathbf{A}$. For example,

$$\mathbf{A} := \begin{pmatrix} 10 & 40 \\ 20 & 50 \\ 30 & 60 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{A}^\top = \begin{pmatrix} 10 & 20 & 30 \\ 40 & 50 & 60 \end{pmatrix} \tag{3.5}$$

and

$$\mathbf{B} := \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{B}^\top = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

A square matrix $\mathbf{A}$ is called **symmetric** if $\mathbf{A}^\top = \mathbf{A}$, or, equivalently, $a_{nk} = a_{kn}$ for every k and n. Note that $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are always well-defined and symmetric.

Fact 3.2.4 For conformable matrices $\mathbf{A}$ and $\mathbf{B}$, transposition satisfies

(i) $(\mathbf{A}^\top)^\top = \mathbf{A}$,

(ii) $(\mathbf{A}\mathbf{B})^\top = \mathbf{B}^\top\mathbf{A}^\top$,

(iii) $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$, and

(iv) $(c\mathbf{A})^\top = c\mathbf{A}^\top$ for any constant c.

Fact 3.2.5 For each square matrix $\mathbf{A}$, we have

(i) $\text{trace}(\mathbf{A}) = \text{trace}(\mathbf{A}^\top)$ and

(ii) $\det(\mathbf{A}^\top) = \det(\mathbf{A})$.

(iii) If $\mathbf{A}$ is nonsingular, then so is $\mathbf{A}^\top$, and its inverse is $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$.

Note that if $\mathbf{a}$ and $\mathbf{b}$ are $N \times 1$ vectors, then the matrix product $\mathbf{a}^\top\mathbf{b} = \mathbf{b}^\top\mathbf{a}$ is equal to $\sum_{n=1}^N a_n b_n$, which is the same as the inner product $\langle \mathbf{a}, \mathbf{b} \rangle$. In what follows we'll often work with column vectors and use matrix product rather than inner product notation.

3.2.3 Eigenvalues and Eigenvectors

Let $\mathbf{A}$ be $N \times N$. As in § 3.1.2, think of $\mathbf{A}$ as a linear map, so that $\mathbf{A}\mathbf{x}$ is the image of $\mathbf{x}$ under $\mathbf{A}$. In general, $\mathbf{A}$ will map $\mathbf{x}$ to some arbitrary new location but sometimes $\mathbf{x}$ will only be *scaled*. That is,

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{3.6}$$

for some scalar λ. If $\mathbf{x}$ and λ satisfy (3.6) and $\mathbf{x}$ is nonzero, then $\mathbf{x}$ is called an **eigenvector** of $\mathbf{A}$, λ is called an **eigenvalue**, and $(\mathbf{x}, \lambda)$ is called an **eigenpair**.

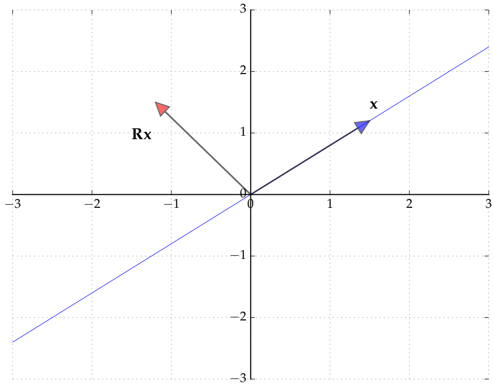

Figure 3.1 **R** rotates points by 90°

Example 3.2.1 If

$$\mathbf{A} := \begin{pmatrix} 1 & -1 \\ 3 & 5 \end{pmatrix}, \quad \mathbf{x} := \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad \lambda := 2$$

then $(\mathbf{x}, \lambda)$ is an eigenpair of $\mathbf{A}$, since $\mathbf{x} \neq \mathbf{0}$ and

$$\mathbf{A}\mathbf{x} = \begin{pmatrix} 1 & -1 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \lambda \mathbf{x}$$

Example 3.2.2 If $\mathbf{I}$ is the $N \times N$ identity then $(\mathbf{x}, 1)$ is an eigenpair of $\mathbf{I}$ for every nonzero $\mathbf{x} \in \mathbb{R}^N$.

Consider now the matrix

$$\mathbf{R} := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

that rotates any point counterclockwise by 90°, as shown in figure 3.1. This rotation causes the scaling in (3.6) to fail for any $\lambda \in \mathbb{R}$ and nonzero $\mathbf{x} \in \mathbb{R}^2$. If, however, we admit the possibility that λ and the elements of $\mathbf{x}$ can be complex, then (3.6) can hold. For example, direct calculation confirms that $\lambda = i$ and $\mathbf{x} = (1, -i)^\mathsf{T}$ is an eigenpair for $\mathbf{R}$. It turns out that contemplation of complex eigenpairs is useful. Here eigenpairs are taken to be complex-valued unless explicitly stated to be real.

Fact 3.2.6 Given $N \times N$ matrix $\mathbf{A}$, the scalar λ is an eigenvalue of $\mathbf{A}$ if and only if

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

Here $\mathbf{I}$ is the $N \times N$ identity. Exercise 3.5.17 asks you to confirm this important fact. For the 2×2 matrix in (3.4), the rule for the 2×2 determinant in (3.4), fact 3.2.6 and a little bit of algebra imply that its eigenvalues are given by the two roots of the polynomial expression

$$\lambda^2 - (a + d)\lambda + (ad - bc) = 0$$

More generally, given any $N \times N$ matrix $\mathbf{A}$, it can be shown via the fundamental theorem of algebra that there exist complex numbers $\lambda_1, \ldots, \lambda_N$, not necessarily distinct, such that

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \prod_{n=1}^{N} (\lambda_n - \lambda) \tag{3.7}$$

It is clear that each λ_n satisfies $\det(\mathbf{A} - \lambda_n \mathbf{I}) = 0$ and hence is an eigenvalue of $\mathbf{A}$. In particular, $\lambda_1, \ldots, \lambda_N$ is the set of eigenvalues of $\mathbf{A}$, although it's worth repeating that these numbers are not necessarily distinct.

Fact 3.2.7 Let $\mathbf{A}$ be $N \times N$ and let $\lambda_1, \ldots, \lambda_N$ be the eigenvalues defined in (3.7). The following statements are true:

(i) $\det \mathbf{A} = \prod_{n=1}^{N} \lambda_n$.

(ii) $\operatorname{trace} \mathbf{A} = \sum_{n=1}^{N} \lambda_n$.

(iii) If $\mathbf{A}$ is symmetric, then $\lambda_n \in \mathbb{R}$ for all n.

(iv) If $\mathbf{A}$ is nonsingular, then the eigenvalues of $\mathbf{A}^{-1}$ are $1/\lambda_1, \ldots, 1/\lambda_N$.

(v) If $\mathbf{A}$ is triangular, then its eigenvalues coincide with the elements on the principle diagonal.

It is immediate from (i) that $\mathbf{A}$ is nonsingular $\iff$ all its eigenvalues are nonzero.

3.2.4 Quadratic Forms

In statistics and econometrics we often encounter quadratic expressions. In general, given symmetric $N \times N$ matrix $\mathbf{A}$, the **quadratic function** or **quadratic form** on $\mathbb{R}^N$ associated with $\mathbf{A}$ is the map Q defined by

$$Q(\mathbf{x}) := \mathbf{x}^\top \mathbf{A} \mathbf{x} = \sum_{j=1}^{N} \sum_{i=1}^{N} a_{ij} x_i x_j$$

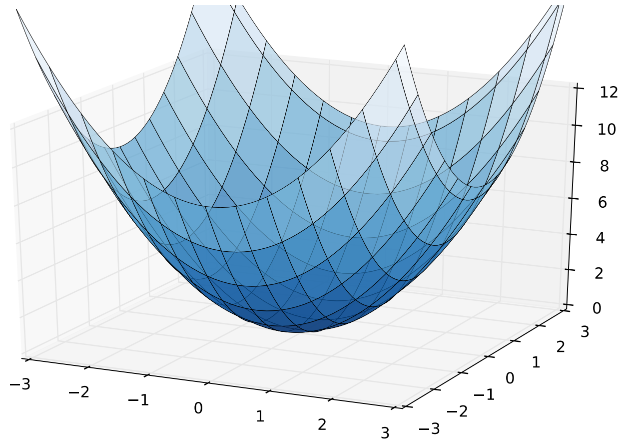

Figure 3.2 Quadratic function $Q(\mathbf{x}) = x_1^2 + x_2^2$

To give a simple illustration, let $N = 2$ and let $\mathbf{A}$ be the identity matrix $\mathbf{I}$. In this case,

$$Q(\mathbf{x}) = \|\mathbf{x}\|^2 = x_1^2 + x_2^2$$

A 3D graph of this function is shown in figure 3.2.

One thing you'll notice about this function is that its graph lies everywhere above zero, or $Q(\mathbf{x}) \geqslant 0$. In fact we know that $\|\mathbf{x}\|^2$ is nonnegative and will be zero only when $\mathbf{x} = \mathbf{0}$. Hence the graph touches zero only at the point $\mathbf{x} = \mathbf{0}$. Many other choices of $\mathbf{A}$ yield a quadratic form with this property. Such $\mathbf{A}$ are said to be positive definite. More generally, an $N \times N$ symmetric matrix $\mathbf{A}$ is called

- **nonnegative definite** if $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geqslant 0$ for all $\mathbf{x} \in \mathbb{R}^N$,
- **positive definite** if $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^N$ with $\mathbf{x} \neq \mathbf{0}$,
- **nonpositive definite** if $\mathbf{x}^\top \mathbf{A} \mathbf{x} \leqslant 0$ for all $\mathbf{x} \in \mathbb{R}^N$, and
- **negative definite** if $\mathbf{x}^\top \mathbf{A} \mathbf{x} < 0$ for all $\mathbf{x} \in \mathbb{R}^N$ with $\mathbf{x} \neq \mathbf{0}$.

If $\mathbf{A}$ fits none of these categories, then $\mathbf{A}$ is called **indefinite**. Figure 3.3 shows the graph of a negative definite quadratic function. Now the function is hill-shaped, and $\mathbf{0}$ is the unique global maximum. Figure 3.4 shows an indefinite form.

The easiest case for detecting definiteness is when the matrix $\mathbf{A}$ is diagonal, since

$$\mathbf{A} = \operatorname{diag}(d_1, \ldots, d_N) \quad \text{implies} \quad Q(\mathbf{x}) = d_1 x_1^2 + \cdots + d_N x_N^2$$

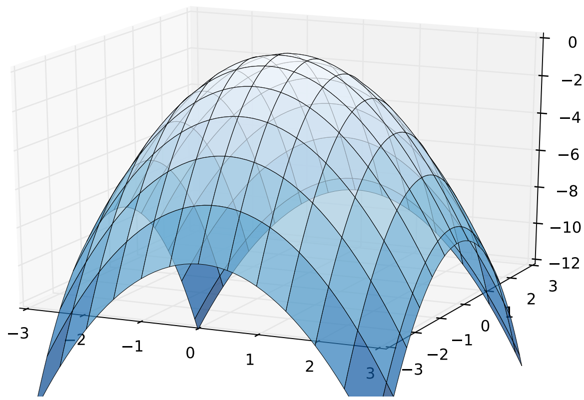

Figure 3.3 Quadratic function $Q(\mathbf{x}) = -x_1^2 - x_2^2$

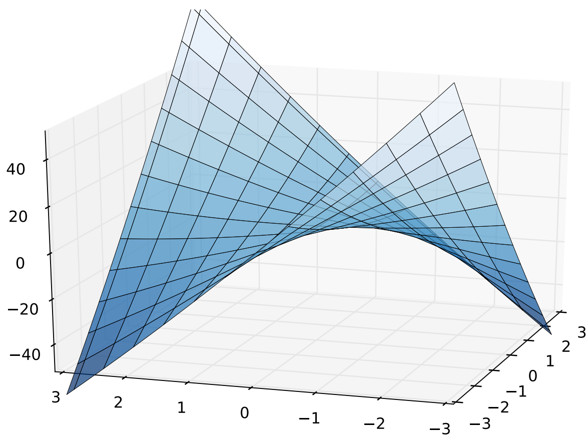

Figure 3.4 Quadratic function $Q(\mathbf{x}) = x_1^2/2 + 8x_1x_2 + x_2^2/2$

From the right-hand expression we see that a diagonal matrix is positive definite if and only if all diagonal elements are positive. Analogous statements are true for nonnegative, nonpositive and negative definite matrices. The next fact generalizes this idea and is proved in §3.3.4.

Fact 3.2.8 Let $\mathbf{A}$ be any symmetric matrix. $\mathbf{A}$ is

(i) positive definite if and only if its eigenvalues are all positive,

(ii) negative definite if and only if its eigenvalues are all negative,

and similarly for nonpositive and nonnegative definite.

It follows from fact 3.2.8 that

Fact 3.2.9 If $\mathbf{A}$ is positive definite, then $\mathbf{A}$ is nonsingular, with $\det \mathbf{A} > 0$.

Finally, here's a necessary (but not sufficient) condition for each kind of definiteness.

Fact 3.2.10 If $\mathbf{A}$ is positive definite, then each element a_{nn} on the principal diagonal is positive, and the same for nonnegative, nonpositive and negative.

3.3 Projection and Decomposition

This section collects some essential results on projection and decomposition of matrices. The projection material takes our abstract projection theory from §2.2 and translates it into the more concrete language of matrices.

3.3.1 Projection Matrices

As stated in theorem 2.2.2 on page 31, given any subspace S, the corresponding projection $\mathbf{P} = \operatorname{proj} S$ is a linear map from $\mathbb{R}^N$ to $\mathbb{R}^N$. In view of theorem 3.1.1 on page 48, it follows that there exists an $N \times N$ matrix $\hat{\mathbf{P}}$ such that $\mathbf{P}\mathbf{x} = \hat{\mathbf{P}}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^N$. In fact we've anticipated this in the notation $\mathbf{P}$, and from now on $\mathbf{P}$ will also represent the corresponding matrix. But what does this matrix look like?

Theorem 3.3.1 *Let S be a subspace of $\mathbb{R}^N$. If $\mathbf{P} = \operatorname{proj} S$, then*

$$\mathbf{P} = \mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top \tag{3.8}$$

for every matrix $\mathbf{B}$ such that the columns of $\mathbf{B}$ form a basis of S.

This result generalizes fact 2.2.6 on page 32, which pertains to projection onto the span of an orthonormal basis. The expression in (3.8) is more complex, but at the same time it applies to any basis, orthonormal or otherwise. We further explore the connection between the two results in §3.3.3. For the proof of theorem 3.3.1 see exercise 3.5.30 and its solution.

The construction of $\mathbf{P}$ in (3.8) implicitly assumes that $\mathbf{B}^\top \mathbf{B}$ is nonsingular. This is justified because $\mathbf{B}$ has full column rank (see ex. 3.5.28). As usual, we let $\mathbf{M} = \mathbf{I} - \mathbf{P}$ denote the residual projection (see page 32).

Example 3.3.1 Recall example 2.2.1 on page 30, where we found that the projection of $\mathbf{y} \in \mathbb{R}^N$ onto span$\{\mathbf{1}\}$ is $\bar{y}\mathbf{1}$. We can get this from (3.8) as well. Since $\mathbf{1}$ is a basis for span$\{\mathbf{1}\}$, we have

$$\mathbf{P} = \text{proj span}\{\mathbf{1}\} \implies \mathbf{P} = \mathbf{1}(\mathbf{1}^\top \mathbf{1})^{-1}\mathbf{1}^\top = \frac{1}{N}\mathbf{1}\mathbf{1}^\top \tag{3.9}$$

This leads us back to $\mathbf{P}\mathbf{y} = \bar{y}\mathbf{1}$, as expected. The corresponding residual projection is

$$\mathbf{M}_c = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \tag{3.10}$$

The reason for the subscripted c was discussed in example 2.2.3 on page 33.

Fact 3.3.1 In the setting of theorem 3.3.1, we have

(i) $\mathbf{M}\mathbf{B} = \mathbf{0}$

(ii) $\mathbf{P}\mathbf{B} = \mathbf{B}$

For example, it's easy to see that $\mathbf{M}_c$ in (3.10) maps $\mathbf{1}$ to $\mathbf{0}$. Exercise 3.5.31 asks you to prove fact 3.3.1.

A square matrix $\mathbf{A}$ is called **idempotent** if $\mathbf{A}\mathbf{A} = \mathbf{A}$.

Fact 3.3.2 Both $\mathbf{P}$ and $\mathbf{M}$ are symmetric and idempotent.

Idempotence of $\mathbf{P}$ and $\mathbf{M}$ can be checked by direct calculation. A better understanding is obtained by reflecting on the fact that projecting onto a subspace twice is the same as projecting once. After one projection the vector is already in the subspace. See fact 2.2.8 on page 33.

Fact 3.3.3 If $\mathbf{A}$ is any idempotent matrix, then rank $\mathbf{A}$ = trace $\mathbf{A}$.

For orthogonal projections, we can say more:

Fact 3.3.4 Let S be a linear subspace of $\mathbb{R}^N$. If $\mathbf{P} = \text{proj}\, S$ and $\mathbf{M}$ is the residual projection, then

(i) $\operatorname{rank} \mathbf{P} = \operatorname{trace} \mathbf{P} = \dim S$ and

(ii) $\operatorname{rank} \mathbf{M} = \operatorname{trace} \mathbf{M} = N - \dim S$.

To see why $\operatorname{rank} \mathbf{P} = \dim S$, recall that the rank of a linear map is the dimension of its range. When $\mathbf{P} = \operatorname{proj} S$, the range of $\mathbf{P}$ is exactly S.

To show that $\operatorname{trace} \mathbf{P} = \dim S$ also holds, we can appeal to fact 3.3.3. There's also a nice direct proof. See exercise 3.5.27 and its solution. Once we know that $\operatorname{trace} \mathbf{P} = \dim S$, it's clear that $\operatorname{trace} \mathbf{M} = N - \dim S$ because

$$\operatorname{trace} \mathbf{M} = \operatorname{trace}(\mathbf{I} - \mathbf{P}) = \operatorname{trace} \mathbf{I} - \operatorname{trace} \mathbf{P} = N - \dim S$$

3.3.2 Overdetermined Systems of Equations

In §3.1.3 we talked about solving equations of the form $\mathbf{A}\mathbf{x} = \mathbf{b}$ when $\mathbf{A}$ is square. In statistics and econometrics, we often work the case where $\mathbf{A}$ is $N \times K$ and $K \leqslant N$. When the inequality is strict, the system of equations is said to be **overdetermined**.

Consider the problem of whether or not there exists a vector $\mathbf{x}$ satisfying $\mathbf{A}\mathbf{x} = \mathbf{b}$ in this setting. Intuitively, when the number of equations is larger than the number of unknowns ($N > K$) we may not be able to find an $\mathbf{x}$ that satisfies all N equations. There are several equivalent ways to formalize this intuition. The linear map $T \colon \mathbb{R}^K \to \mathbb{R}^N$ corresponding to $\mathbf{A}$ is $T\mathbf{x} = \mathbf{A}\mathbf{x}$ (see 3.1.2). We know the following statements to be equivalent:

(i) there exists an $\mathbf{x} \in \mathbb{R}^K$ with $\mathbf{A}\mathbf{x} = \mathbf{b}$.

(ii) $\mathbf{b} \in \operatorname{colspace} \mathbf{A}$.

(iii) $\mathbf{b} \in \operatorname{rng} T$.

We also know from theorem 2.1.8 on page 26 that when $K < N$, the function T cannot be onto, and hence it's possible that $\mathbf{b}$ lies outside the range of T.

In fact we can say more than this. When $K < N$, the scenario $\mathbf{b} \in \operatorname{colspace} \mathbf{A}$ is in some sense very rare. The reason is that the dimension of $\operatorname{colspace} \mathbf{A}$, which is precisely the rank of $\mathbf{A}$, is less than or equal to K (see §3.1.2) and hence strictly less than N. There is a sense in which K-dimensional subspaces of $\mathbb{R}^N$ are negligible, and hence the "chance" of $\mathbf{b}$ happening to lie in this subspace is tiny.[1]

As a result the standard approach is to admit that an exact solution may not exist, and to focus instead on finding a $\mathbf{x} \in \mathbb{R}^K$ such that $\mathbf{A}\mathbf{x}$ is as close to $\mathbf{b}$ as possible. It's

1. Formally, K dimensional subspaces have measure zero in $\mathbb{R}^N$ whenever $K < N$. Hence every probability measure that is absolutely continuous with respect to Lebesgue measure puts zero mass on such a set. Less formally, consider the case where $N = 3$ and $K = 2$. Then $\operatorname{colspace} \mathbf{A}$ forms at most a 2-dimensional plane in $\mathbb{R}^3$. Intuitively, this set has no volume in $\mathbb{R}^3$ because planes have no "thickness." Similarly, while we might visualize $\mathbf{b}$ as a dot in $\mathbb{R}^3$, as a point it is in fact infinitesimally small. Hence the chance of a randomly chosen $\mathbf{b}$ lying in $\operatorname{colspace} \mathbf{A}$ is zero.

natural that “close to" is defined in terms of ordinary Euclidean norm, which leads us to the minimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^K} \|\mathbf{b} - \mathbf{A}\mathbf{x}\| \tag{3.11}$$

This is called a **least squares problem** because solving (3.11) is the same as minimizing $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2$ with respect to $\mathbf{x}$, and the squared norm is, by definition, a sum of squares. Assuming as before that $\mathbf{A}$ is $N \times K$ with $K \leqslant N$ and $\mathbf{b}$ is $N \times 1$, we can use the orthogonal projection theorem to solve (3.11) as follows:

Theorem 3.3.2 *If* $\mathbf{A}$ *has full column rank, then* (3.11) *has the unique solution*

$$\hat{\mathbf{x}} := (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b} \tag{3.12}$$

Proof. Let $\mathbf{A}$ and $\mathbf{b}$ be as in the statement of the theorem. Let $\hat{\mathbf{x}}$ be as in (3.12) and let $S := \operatorname{colspace} \mathbf{A}$. By the full column rank assumption, the columns of $\mathbf{A}$ form a basis for S. Hence, applying theorem 3.3.1, the orthogonal projection of $\mathbf{b}$ onto S is

$$\mathbf{P}\mathbf{b} := \mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b} = \mathbf{A}\hat{\mathbf{x}} \tag{3.13}$$

Moreover, since the orthogonal projection theorem gives a unique minimizer in terms of the closest point in S to $\mathbf{b}$, we must have

$$\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}\| < \|\mathbf{b} - \mathbf{y}\| \quad \text{for all } \mathbf{y} \in S, \ \mathbf{y} \neq \mathbf{A}\hat{\mathbf{x}} \tag{3.14}$$

Pick any $\mathbf{x} \in \mathbb{R}^K$ such that $\mathbf{x} \neq \hat{\mathbf{x}}$. By the definition of S we have $\mathbf{A}\mathbf{x} \in S$. In addition, since $\mathbf{x} \neq \hat{\mathbf{x}}$, and since $\mathbf{A}$ has full column rank, it must be that $\mathbf{A}\mathbf{x} \neq \mathbf{A}\hat{\mathbf{x}}$ (ex. 3.5.4). Hence

$$\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}\| < \|\mathbf{b} - \mathbf{A}\mathbf{x}\| \quad \text{for all } \mathbf{x} \in \mathbb{R}^K, \ \mathbf{x} \neq \hat{\mathbf{x}}$$

In other words, $\hat{\mathbf{x}}$ is the unique solution to (3.11). □

In the expression for $\hat{\mathbf{x}}$ in (3.12), the matrix $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}$ is called the **pseudoinverse** of $\mathbf{A}$. If $K = N$—that is, if $\mathbf{A}$ is in fact square—then, under our full rank assumption, the pseudoinverse reduces to the inverse $\mathbf{A}^{-1}$, and the least squares solution $\hat{\mathbf{x}}$ in (3.12) reduces to the expression given in (3.3) on page 51.

What happens if we drop the assumption that the columns of $\mathbf{A}$ are linearly independent? The set $\operatorname{colspace} \mathbf{A}$ is still a linear subspace, and the orthogonal projection theorem still gives us a closest point $\mathbf{P}\mathbf{b}$ to $\mathbf{b}$ in $\operatorname{colspace} \mathbf{A}$. Since $\mathbf{P}\mathbf{b} \in \operatorname{colspace} \mathbf{A}$, there still exists a vector $\hat{\mathbf{x}}$ such that $\mathbf{P}\mathbf{b} = \mathbf{A}\hat{\mathbf{x}}$. The problem is that now there exists an infinity of such vectors. Exercise 3.5.34 asks you to prove this.

3.3.3 QR Decomposition

The **QR decomposition** of a given matrix $\mathbf{A}$ is a product of the form $\mathbf{QR}$, where the first matrix has orthonormal columns and the second is upper triangular. This factorization has many applications, including least squares problems and the computation of eigenvalues. The decomposition is based on Gram–Schmidt orthogonalization.

Theorem 3.3.3 *If $\mathbf{A}$ is an $N \times K$ matrix with full column rank, then there exists a factorization $\mathbf{A} = \mathbf{QR}$ where*

(i) $\mathbf{R}$ is $K \times K$, upper triangular and nonsingular, and

(ii) $\mathbf{Q}$ is $N \times K$, with orthonormal columns.

Proof. Let $\mathbf{A}$ be as stated, and let $\mathbf{a}_k := \operatorname{col}_k \mathbf{A}$. Theorem 2.2.3 gives us existence of an orthonormal set $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ such that the span of $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ equals that of $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ for $k = 1, \ldots, K$. In particular, $\mathbf{a}_k$ is in the span of $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, and hence, appealing to fact 2.2.3 on page 28, we can write

$$
\begin{aligned}
\mathbf{a}_1 &= (\mathbf{a}_1^\top \mathbf{u}_1)\mathbf{u}_1 \\
\mathbf{a}_2 &= (\mathbf{a}_2^\top \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{a}_2^\top \mathbf{u}_2)\mathbf{u}_2 \\
\mathbf{a}_3 &= (\mathbf{a}_3^\top \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{a}_3^\top \mathbf{u}_2)\mathbf{u}_2 + (\mathbf{a}_3^\top \mathbf{u}_3)\mathbf{u}_3
\end{aligned}
$$

and so on. Sticking to the 3×3 case to simplify expressions, we can stack these equations horizontally to get

$$
\begin{pmatrix} | & | & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \\ | & | & | \end{pmatrix}
=
\begin{pmatrix} | & | & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ | & | & | \end{pmatrix}
\begin{pmatrix} (\mathbf{a}_1^\top \mathbf{u}_1) & (\mathbf{a}_2^\top \mathbf{u}_1) & (\mathbf{a}_3^\top \mathbf{u}_1) \\ 0 & (\mathbf{a}_2^\top \mathbf{u}_2) & (\mathbf{a}_3^\top \mathbf{u}_2) \\ 0 & 0 & (\mathbf{a}_3^\top \mathbf{u}_3) \end{pmatrix}
$$

or $\mathbf{A} = \mathbf{QR}$. This is our QR decomposition.

It remains to show that $\mathbf{R}$ is invertible. This will be so if each term $\mathbf{a}_k^\top \mathbf{u}_k$ is nonzero, since the determinant is the product of these elements (see fact 3.2.7 on page 57). Suppose, to the contrary, that $\mathbf{a}_k^\top \mathbf{u}_k = 0$ for some k. Then $\mathbf{a}_k$ lies in the span of $\{\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}\}$, which, by construction, agrees with the span of $\{\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}\}$. This contradicts linear independence of the columns of $\mathbf{A}$. □

Given the decomposition $\mathbf{A} = \mathbf{QR}$, the least squares solution $\hat{\mathbf{x}}$ defined in (3.12) can also be written as $\hat{\mathbf{x}} = \mathbf{R}^{-1}\mathbf{Q}^\top \mathbf{b}$ (ex. 3.5.32). Premultiplying by $\mathbf{R}$ converts this expression to $\mathbf{R}\hat{\mathbf{x}} = \mathbf{Q}^\top \mathbf{b}$, which is easy to solve because $\mathbf{R}$ is triangular (see §3.2.1).

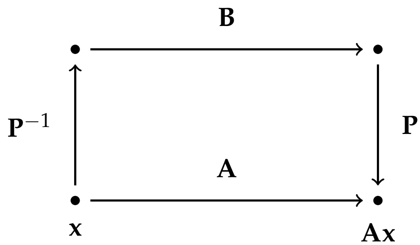

Figure 3.5 **A** is similar to **B**

3.3.4 Diagonalization and Spectral Theory

One important concept in dynamic systems and related fields is topological conjugacy. If $f\colon A \to A$ and $g\colon B \to B$, then g is said to be **topologically conjugate** to f whenever there exists a continuous bijection $\tau\colon B \to A$ such that $f = \tau \circ g \circ \tau^{-1}$. The idea is that the action of f can be replicated by transporting a point into the domain of g, applying g, and then transporting it back. This can be beneficial if g is somehow simpler than f.

In the case of linear maps—that is, matrices—it is natural to study conjugacy in a setting where the bijection is also required to be linear. In linear algebra this is called *similarity*. In particular, a square matrix $\mathbf{A}$ is said to be **similar** to another matrix $\mathbf{B}$ if there exists an invertible matrix $\mathbf{P}$ such that $\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}^{-1}$. Figure 3.5 shows the conjugate relationship of the two matrices when thought of as maps.

The next fact is a fun exercise.

Fact 3.3.5 If $\mathbf{A}$ is similar to $\mathbf{B}$, then $\mathbf{A}^t$ is similar to $\mathbf{B}^t$ for all $t \in \mathbb{N}$.

As discussed above, similarity of $\mathbf{A}$ to a given matrix $\mathbf{B}$ is most useful when $\mathbf{B}$ is somehow simpler than $\mathbf{A}$, or more amenable to a given operation. The simplest kind of matrices we work with are diagonal matrices, so similarity to a diagonal matrix is particularly desirable. If $\mathbf{A}$ is similar to a diagonal matrix, then $\mathbf{A}$ is called **diagonalizable**.

Example 3.3.2 Suppose that we want to calculate $\mathbf{A}^t$ for some given $t \in \mathbb{N}$. If $\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$ for some $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$, then by fact 3.3.5 and fact 3.2.2 on page 54, we have $\mathbf{A}^t = \mathbf{P}\operatorname{diag}(\lambda_1^t, \ldots, \lambda_N^t)\mathbf{P}^{-1}$. Aside from mapping backward and forward with $\mathbf{P}$, the only effort required to take the tth power of $\mathbf{A}$ is to take the tth power of N scalars.

If $\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$ where $\boldsymbol{\Lambda}$ is diagonal, it follows immediately that the elements on the principal diagonal of $\boldsymbol{\Lambda}$ are the eigenvalues of $\mathbf{A}$, and that the columns of $\mathbf{P}$ are eigenvectors. To see this, observe that $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ implies $\mathbf{AP} = \mathbf{P}\mathbf{\Lambda}$. Equating the nth column on each side gives $\mathbf{A}\mathbf{p}_n = \lambda_n \mathbf{p}_n$, where $\mathbf{p}_n := \operatorname{col}_n \mathbf{P}$. Finally, $\mathbf{p}_n$ is not the zero vector because otherwise $\mathbf{P}$ would not be invertible. Let's summarize:

Fact 3.3.6 If $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ for some $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$, then $(\operatorname{col}_n \mathbf{P}, \lambda_n)$ is an eigenpair of $\mathbf{A}$ for each n.

But when is $\mathbf{A}$ diagonalizable? The only thing we need to make the expression $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ work is for $\mathbf{P}$ to be invertible. This explains the next fact.

Fact 3.3.7 An $N \times N$ matrix $\mathbf{A}$ is diagonalizable if and only if it has N linearly independent eigenvectors.[2]

If $\mathbf{A}$ does have N linearly independent eigenvectors, then we are in good shape—diagonalization is possible. But can we do any better? The matrix $\mathbf{\Lambda}$ cannot really be simplified, but in some cases $\mathbf{P}$ can be. In particular, things are even nicer if $\mathbf{P}$ has orthonormal columns—that is, its columns form an orthonormal set, and hence an orthonormal basis of $\mathbb{R}^N$. These kinds of matrices are called **orthogonal matrices**. Here are some nice properties of such matrices.

Fact 3.3.8 If $\mathbf{Q}$ and $\mathbf{P}$ are $N \times N$ orthogonal matrices, then

(i) $\mathbf{Q}^\top$ is orthogonal and $\mathbf{Q}^{-1} = \mathbf{Q}^\top$,

(ii) $\mathbf{QP}$ is orthogonal, and

(iii) $\det \mathbf{Q} \in \{-1, 1\}$.

The first result tells us is that if $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ and $\mathbf{Q}$ has orthonormal columns, then $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$. It's easy to see from this expression that $\mathbf{A}$ must be symmetric, so if we hope to diagonalize with this extra orthonormal property, then we shouldn't look beyond symmetric matrices. The following fundamental theorem tells us that this form of diagonalization is possible precisely when $\mathbf{A}$ is symmetric.

Theorem 3.3.4 *If $\mathbf{A}$ is symmetric, then $\mathbf{A}$ can be diagonalized as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is the diagonal matrix formed from the eigenvalues of $\mathbf{A}$.*

This is one version of the **spectral decomposition theorem**. See, for example, §10.3 of Jänich (1994). $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ is called the **symmetric eigenvalue decomposition** of $\mathbf{A}$. It's not hard to see that its action on a given $N \times 1$ vector $\mathbf{x}$ can be written as

$$\mathbf{A}\mathbf{x} = \sum_{n=1}^{N} \lambda_n (\mathbf{u}_n^\top \mathbf{x}) \mathbf{u}_n$$

2. If we permit the eigenvectors to be complex, then the requirement is that they form a basis of $\mathbb{C}^N$, the set of complex N-vectors. In this case there is a sense in which almost all matrices are diagonalizable.

where λ_n is the nth eigenvalue of $\mathbf{A}$ and $\mathbf{u}_n = \text{col}_n\, \mathbf{Q}$. Compare with $\mathbf{x} = \sum_{n=1}^{N} (\mathbf{u}_n^\top \mathbf{x}) \mathbf{u}_n$, which is true by fact 2.2.3 on page 28.

One nice application of the spectral theorem is a proof of fact 3.2.8 on page 60, which states among other things that symmetric matrix $\mathbf{A}$ is positive definite if and only if its eigenvalues are all positive. Exercise 3.5.23 and its solution step you through the arguments.

Fact 3.3.9 If $\mathbf{A}$ is nonnegative definite, then $\sqrt{\mathbf{A}}$ exists and equals $\mathbf{Q}\sqrt{\mathbf{\Lambda}}\mathbf{Q}^\top$. The matrix $\sqrt{\mathbf{\Lambda}}$ is given by $\text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_N})$.

Here $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ is the spectral decomposition of $\mathbf{A}$. To check that $\mathbf{Q}\sqrt{\mathbf{\Lambda}}\mathbf{Q}^\top$ is a square root just multiply. $\sqrt{\mathbf{\Lambda}}$ exists because $\mathbf{A}$ is nonnegative definite, and hence its eigenvalues are nonnegative (see fact 3.2.8 and fact 3.2.2).

By combining our results on spectral decomposition with the QR decomposition discussed in §3.3.3, we can prove the following well-known fact:

Fact 3.3.10 If $\mathbf{A}$ is positive definite, then there exists a nonsingular, upper triangular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{R}^\top \mathbf{R}$.

This decomposition is called the **Cholesky decomposition**. The proof is obtained by writing

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

and applying the QR decomposition to $\sqrt{\mathbf{\Lambda}}\mathbf{Q}^\top$. This allows us to write $\sqrt{\mathbf{\Lambda}}\mathbf{Q}^\top = \tilde{\mathbf{Q}}\mathbf{R}$, where $\mathbf{R}$ is nonsingular and upper triangular, and $\tilde{\mathbf{Q}}$ has orthonormal columns. Because the columns of $\tilde{\mathbf{Q}}$ are orthonormal,

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

Our decomposition has the properties stated in fact 3.3.10.

3.3.5 Norms and Continuity

Given vector sequence $\{\mathbf{x}_n\}$ in $\mathbb{R}^K$ and any point $\mathbf{x} \in \mathbb{R}^K$, we say that $\{\mathbf{x}_n\}$ **converges** to $\mathbf{x}$ and write $\mathbf{x}_n \to \mathbf{x}$ if, for any $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that $\|\mathbf{x}_n - \mathbf{x}\| < \epsilon$ whenever $n \geqslant N$. Another way to say this is that the real-valued sequence $z_n := \|\mathbf{x}_n - \mathbf{x}\|$ converges to zero in $\mathbb{R}$ as $n \to \infty$.

Fact 3.3.11 The following results hold:

(i) If $\mathbf{x}_n \to \mathbf{x}$ and $\mathbf{y}_n \to \mathbf{y}$, then $\mathbf{x}_n + \mathbf{y}_n \to \mathbf{x} + \mathbf{y}$.

(ii) If $\mathbf{x}_n \to \mathbf{x}$ and $\alpha \in \mathbb{R}$, then $\alpha\mathbf{x}_n \to \alpha\mathbf{x}$.

(iii) $\mathbf{x}_n \to \mathbf{x}$ if and only if $\mathbf{a}^\top \mathbf{x}_n \to \mathbf{a}^\top \mathbf{x}$ for all $\mathbf{a} \in \mathbb{R}^K$.

For econometrics it is helpful if we extend the notion of convergence to matrices. We can do this in a parallel fashion by defining norms over matrices. The **matrix norm** of $N \times K$ matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\| := \max \left\{ \|\mathbf{A}\mathbf{x}\| \,:\, \mathbf{x} \in \mathbb{R}^K,\ \|\mathbf{x}\| = 1 \right\} \tag{3.15}$$

In this expression there are two different norms in play: The left-hand side is a matrix norm, while on the right we have the ordinary vector norm.

Given a sequence of $N \times K$ matrices $\mathbf{A}_n$ and an $N \times K$ matrix $\mathbf{A}$, we say that $\mathbf{A}_n$ **converges** to $\mathbf{A}$ if the matrix norm deviation $\|\mathbf{A}_n - \mathbf{A}\|$ converges to zero in $\mathbb{R}$.

While the value in (3.15) is not particularly easy to solve for in general, the definition is entirely standard, and one can show that the matrix norm behaves like the vector norm in many ways. For example,

Fact 3.3.12 For any conformable matrices $\mathbf{A}$ and $\mathbf{B}$, the matrix norm satisfies

(i) $\|\mathbf{A}\| \geqslant 0$ and $\|\mathbf{A}\| = 0$ if and only if all entries of $\mathbf{A}$ are zero,

(ii) $\|\alpha \mathbf{A}\| = |\alpha| \|\mathbf{A}\|$ for any scalar α,

(iii) $\|\mathbf{A} + \mathbf{B}\| \leqslant \|\mathbf{A}\| + \|\mathbf{B}\|$, and

(iv) $\|\mathbf{A}\mathbf{B}\| \leqslant \|\mathbf{A}\| \|\mathbf{B}\|$.

Compare these results with those of fact 2.1.2 on page 13.

Fact 3.3.13 For any $J \times K$ matrix $\mathbf{A}$ with elements a_{jk}, we have

$$\|\mathbf{A}\| \leqslant \sqrt{JK} \max_{jk} |a_{jk}|$$

This bound is handy. For example, it tells us that if every element of $\mathbf{A}$ is close to zero then $\|\mathbf{A}\|$ must also be close to zero.

3.3.5.1 Neumann Series

Let's look at an important result in analysis that uses matrix norms. Starting in chapter 7, we will investigate dynamic systems such as $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{b}$, where $\mathbf{x}_t$ represents the values of some variables of interest and $\mathbf{A}$ and $\mathbf{b}$ form the parameters in the law of motion for $\mathbf{x}_t$. One question we might ask in this setting is whether or not there exists a "stationary" vector $\mathbf{x} \in \mathbb{R}^N$, in the sense that $\mathbf{x}_t = \mathbf{x}$ implies $\mathbf{x}_{t+1} = \mathbf{x}$. In other words, we seek an $\mathbf{x} \in \mathbb{R}^N$ that solves the system of equations

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b} \qquad (\mathbf{A} \text{ is } N \times N \text{ and } \mathbf{b} \text{ is } N \times 1) \tag{3.16}$$

We can get some insight by reflecting on the scalar case $x = ax + b$. If $|a| < 1$, then this equation has the unique solution

$$\bar{x} = \frac{b}{1-a} = b \sum_{k=0}^{\infty} a^k$$

The second equality follows from elementary results on geometric series.

It turns out that a similar result is true in $\mathbb{R}^N$ if we replace the condition $|a| < 1$ with an analogous result for matrices based around the matrix norm. We begin with the **Neumann series lemma**, which states the following:

Theorem 3.3.5 *If* $\mathbf{A}$ *is square and* $\|\mathbf{A}^j\| < 1$ *for some* $j \in \mathbb{N}$, *then* $\mathbf{I} - \mathbf{A}$ *is invertible, and*

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{i=0}^{\infty} \mathbf{A}^i \tag{3.17}$$

Here $\mathbf{I}$ is the identity. The equality in (3.17) means that the matrix sum $\sum_{i=0}^{t} \mathbf{A}^i$ converges to the left-hand side in matrix norm as $t \to \infty$. The sum is called the **Neumann series** associated with $\mathbf{A}$. The condition in the theorem ensures that it converges. When the condition holds, (3.16) has the unique solution

$$\bar{\mathbf{x}} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} = \sum_{i=0}^{\infty} \mathbf{A}^i \mathbf{b}$$

How to test the condition in theorem 3.3.5? The most commonly used sufficient condition involves the **spectral radius** of $\mathbf{A}$, which is defined as

$$\varrho(\mathbf{A}) := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } \mathbf{A}\} \tag{3.18}$$

In this definition, $|\lambda|$ is the **modulus** of the possibly complex number λ.[3]

Fact 3.3.14 If $\varrho(\mathbf{A}) < 1$, then $\|\mathbf{A}^j\| < 1$ for some $j \in \mathbb{N}$.

To understand why $\varrho(\mathbf{A}) < 1$ is sufficient for the result in the Neumann series lemma, consider theorem 3.3.5 in this light: The claim is that $\sum_{i=0}^{\infty} \mathbf{A}^i$ is the inverse of $\mathbf{I} - \mathbf{A}$, so $\sum_{i=0}^{t} \mathbf{A}^i(\mathbf{I} - \mathbf{A})$ should be close to $\mathbf{I}$ for large t. Evidently

$$\sum_{i=0}^{t} \mathbf{A}^i(\mathbf{I} - \mathbf{A}) = \sum_{i=0}^{t} \mathbf{A}^i - \sum_{i=0}^{t} \mathbf{A}^{i+1} = \mathbf{I} - \mathbf{A}^{t+1}$$

3. The modulus of $a + ib \in \mathbb{C}$ is $(a^2 + b^2)^{1/2}$. If the imaginary part is zero, this reduces to the usual notion of absolute value.

Hence, for the result to go through, we require $\mathbf{A}^t \to \mathbf{0}$ as $t \to \infty$. The case where $\mathbf{A}$ is diagonalizable gives the clearest insight here, so let's suppose, as in §3.3.4, that $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues $\lambda_1, \ldots, \lambda_N$ of $\mathbf{A}$ on its principal diagonal. As discussed in fact 3.3.5 on page 65, we then have

$$\mathbf{A}^t = \mathbf{P} \begin{pmatrix} \lambda_1^t & 0 & \cdots & 0 \\ 0 & \lambda_2^t & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_N^t \end{pmatrix} \mathbf{P}^{-1}$$

If $\varrho(\mathbf{A}) < 1$, then $|\lambda_n| < 1$ for all n, and hence $\lambda_n^t \to 0$ as $t \to \infty$. It follows that $\mathbf{A}^t \to \mathbf{0}$ as required.

3.4 Further Reading

Good treatments of matrix and linear algebra include Jänich (1994) and Axler (2015).

3.5 Exercises

Ex. 3.5.1 Prove that the inverse of a nonsingular matrix is always nonsingular.

Ex. 3.5.2 Prove the claim in fact 3.2.9 on page 60 that if $\mathbf{A}$ is positive definite, then $\mathbf{A}$ is nonsingular. If you can, prove it without invoking positivity of its eigenvalues.

Ex. 3.5.3 Prove fact 3.2.10 on page 60.

Ex. 3.5.4 Let $\mathbf{A}$ be $N \times K$ and full column rank. Show that $\mathbf{x}, \mathbf{z} \in \mathbb{R}^K$ and $\mathbf{x} \neq \mathbf{z}$ implies $\mathbf{A}\mathbf{x} \neq \mathbf{A}\mathbf{z}$.

Ex. 3.5.5 Prove theorem 3.1.2 on page 51. Before doing so, you might like to review fact 2.1.9 on page 26 and theorem 3.1.1 on page 48.

Ex. 3.5.6 Let $\mathbf{A}$ be square. Assuming existence of the inverse $\mathbf{A}^{-1}$, show that $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$.

Ex. 3.5.7 Show that if $\mathbf{e}_i$ and $\mathbf{e}_j$ are the ith and jth canonical basis vectors of $\mathbb{R}^N$ respectively, and $\mathbf{A}$ is an $N \times N$ matrix, then $\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_j = a_{ij}$, the i, jth element of $\mathbf{A}$.

Ex. 3.5.8 Suppose that $\mathbf{A}$ is $N \times K$, the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution, and $K > N$. Show that the same equation has an infinity of solutions.[4]

4. This is the so-called "underdetermined" case, where the number of equations is less than the number of unknowns. Intuitively, we do not have enough restrictions to pin down values uniquely.

Ex. 3.5.9 Let

$$\mathbf{A} := \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{B} := \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

Show that

(i) $\mathbf{A}$ is nonnegative definite.

(ii) $\mathbf{B}$ is *not* positive definite.

Ex. 3.5.10 Let $\mathbf{A}_1, \ldots, \mathbf{A}_J$ be invertible matrices. Use induction and fact 3.1.5 on page 52 to show that the product of these matrices is invertible, and, in particular, that

$$(\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_J)^{-1} = \mathbf{A}_J^{-1} \cdots \mathbf{A}_2^{-1} \mathbf{A}_1^{-1}$$

Ex. 3.5.11 Show that for any matrix $\mathbf{A}$, the matrix $\mathbf{A}^\top \mathbf{A}$ is well-defined (i.e., multiplication is possible), square, and nonnegative definite.

Ex. 3.5.12 Show that if $\mathbf{A}$ and $\mathbf{B}$ are positive definite and $\mathbf{A} + \mathbf{B}$ is well-defined, then $\mathbf{A} + \mathbf{B}$ is also positive definite.

Ex. 3.5.13 Let $\mathbf{A}$ be $N \times K$. Show that if $\mathbf{A}\mathbf{x} = \mathbf{0}$ for all $K \times 1$ vectors $\mathbf{x}$, then $\mathbf{A} = \mathbf{0}$ (i.e., every element of $\mathbf{A}$ is zero). Show as a corollary that if $\mathbf{A}$ and $\mathbf{B}$ are $N \times K$ and $\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}$ for all $K \times 1$ vectors $\mathbf{x}$, then $\mathbf{A} = \mathbf{B}$.

Ex. 3.5.14 Let $\mathbf{I}$ be the $N \times N$ identity matrix.

(i) Explain why $\mathbf{I}$ is full column rank.

(ii) Show that $\mathbf{I}$ is the inverse of itself.

(iii) Let $\mathbf{A} := \alpha \mathbf{I}$. Give a condition on α such that $\mathbf{A}$ is positive definite.

Ex. 3.5.15 Let $\mathbf{X} := \mathbf{I} - 2\mathbf{u}\mathbf{u}^\top$, where $\mathbf{u}$ is an $N \times 1$ vector with $\|\mathbf{u}\| = 1$. Show that $\mathbf{X}$ is symmetric and $\mathbf{X}\mathbf{X} = \mathbf{I}$.

Ex. 3.5.16 Recall the definition of similarity of matrices, as given in §3.3.4. Let's write $\mathbf{A} \sim \mathbf{B}$ if $\mathbf{A}$ is similar to $\mathbf{B}$. Show that $\sim$ is an **equivalence relation** on the set of $N \times N$ matrices. In particular, show that, for any $N \times N$ matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, we have (i) $\mathbf{A} \sim \mathbf{A}$, (ii) $\mathbf{A} \sim \mathbf{B}$ implies $\mathbf{B} \sim \mathbf{A}$ and (iii) $\mathbf{A} \sim \mathbf{B}$ and $\mathbf{B} \sim \mathbf{C}$ implies $\mathbf{A} \sim \mathbf{C}$.

Ex. 3.5.17 Confirm the claim in fact 3.2.6 on page 56.

Ex. 3.5.18 Show that $\mathbf{A}$ is nonsingular if and only if 0 is *not* an eigenvalue for $\mathbf{A}$.

Ex. 3.5.19 Show that the only nonsingular idempotent matrix is the identity matrix.

Ex. 3.5.20 Let $\mathbf{1}$ be $N \times 1$ and let $\mathbf{P} := \frac{1}{N} \mathbf{1}\mathbf{1}^\top$. Verify that $\mathbf{P}$ is idempotent.

Ex. 3.5.21 Show that, for conformable and suitably invertible matrices $\mathbf{A}$, $\mathbf{U}$ and $\mathbf{W}$, we have

$$(\mathbf{A} + \mathbf{U}\mathbf{W})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{I} + \mathbf{W}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{W}\mathbf{A}^{-1}$$

Ex. 3.5.22 Let $\mathbf{Q}$ be an orthogonal matrix. Show that $\mathbf{Q}^{-1} = \mathbf{Q}^{\top}$ and $\det \mathbf{Q} \in \{-1, 1\}$ both hold.

Ex. 3.5.23 Use theorem 3.3.4 (page 66) to prove the following part of fact 3.2.8: A symmetric matrix $\mathbf{A}$ is positive definite if and only if its eigenvalues are all positive.

Ex. 3.5.24 Show that if $\mathbf{Q}$ is an $N \times N$ orthogonal matrix, then $\mathbf{Q}$ is an isometry on $\mathbb{R}^N$. That is, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, we have $\|\mathbf{Q}\mathbf{x} - \mathbf{Q}\mathbf{y}\| = \|\mathbf{x} - \mathbf{y}\|$.

Ex. 3.5.25 Let $\mathbf{P}$ be square, symmetric and idempotent. Let $S :=$ colspace $\mathbf{P}$. Show that $\mathbf{P} = \operatorname{proj} S$.

Ex. 3.5.26 Consider theorem 3.3.2 on page 63. If $N = K$, what does $\hat{\mathbf{x}}$ reduce to? Interpret.

Ex. 3.5.27 Let S be a linear subspace of $\mathbb{R}^N$ and let $\mathbf{P} = \operatorname{proj} S$. Show that trace $\mathbf{P} = \dim S$ without using the idempotence connection in fact 3.3.3.

Ex. 3.5.28 Show that when $N \times K$ matrix $\mathbf{B}$ is full column rank, the matrix $\mathbf{B}^{\top}\mathbf{B}$ is nonsingular.[5]

Ex. 3.5.29 Let $\mathbf{A}$ be an $N \times N$ matrix.

(i) Show that if $\mathbf{I} - \mathbf{A}$ is idempotent, then $\mathbf{A}$ is idempotent.

(ii) Show that if $\mathbf{A}$ is both symmetric and idempotent, then the matrix $\mathbf{I} - 2\mathbf{A}$ is orthogonal.

Ex. 3.5.30 Prove theorem 3.3.1 on page 60. (This takes a bit of work, of course.)

Ex. 3.5.31 Let $\mathbf{P} = \mathbf{B}(\mathbf{B}^{\top}\mathbf{B})^{-1}\mathbf{B}^{\top}$, as in theorem 3.3.1, and let $\mathbf{M}$ be the residual projection. Show that $\mathbf{M}\mathbf{B} = \mathbf{0}$ using matrix algebra.

Ex. 3.5.32 Let $\mathbf{A}$ be an $N \times K$ matrix with linearly independent columns and QR factorization $\mathbf{A} = \mathbf{Q}\mathbf{R}$ (see §3.3.3). Fix $\mathbf{b} \in \mathbb{R}^N$. Show that $\hat{\mathbf{x}}$ defined in (3.12) can also be written as $\hat{\mathbf{x}} = \mathbf{R}^{-1}\mathbf{Q}^{\top}\mathbf{b}$.

Ex. 3.5.33 Let's prove the Cauchy–Schwarz inequality $|\mathbf{x}^{\top}\mathbf{y}| \leqslant \|\mathbf{x}\|\|\mathbf{y}\|$ from fact 2.1.2 on page 13 via the orthogonal projection theorem. Let $\mathbf{y}$ and $\mathbf{x}$ be nonzero vectors in $\mathbb{R}^N$ (since if either equals zero then the inequality is trivial), and let span$\{\mathbf{x}\}$ be all vectors of the form $\alpha\mathbf{x}$ for $\alpha \in \mathbb{R}$.

5. Hint: In view of fact 3.2.9, it suffices to show that $\mathbf{B}^{\top}\mathbf{B}$ is positive definite.

(i) Letting $\mathbf{P}$ be the orthogonal projection onto span$\{\mathbf{x}\}$, show that

$$\mathbf{P}\mathbf{y} = \frac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}} \mathbf{x}$$

(ii) Using this expression and any relevant properties of orthogonal projections (see theorem 2.2.2 on page 31), confirm the Cauchy–Schwarz inequality.

Ex. 3.5.34 Prove the claim made after theorem 3.3.2 regarding failure of uniqueness without full column rank. In particular, let $\mathbf{A}$ be $N \times K$ with linearly *dependent* columns, and let $\mathbf{P}\mathbf{b}$ be the closest point to $\mathbf{b}$ in colspace $\mathbf{A}$. Prove that there are infinitely many $\mathbf{x} \in \mathbb{R}^K$ such that $\mathbf{P}\mathbf{b} = \mathbf{A}\mathbf{x}$.

Ex. 3.5.35 Let $\mathbf{A}$ be symmetric and idempotent. Show that every eigenvalue of $\mathbf{A}$ is either 0 or 1.

Ex. 3.5.36 Show that if $\mathbf{A}$ is positive definite, then there exists a symmetric matrix $\mathbf{C}$ such that $\mathbf{CAC} = \mathbf{I}$.[6]

3.5.1 Solutions to Selected Exercises

Solution to Ex. 3.5.1. Let $\mathbf{A}$ be a nonsingular matrix. Being nonsingular, $\mathbf{A}$ is invertible, with inverse $\mathbf{A}^{-1}$. By the definition of the inverse, we have $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the identity. This tells us directly that $\mathbf{A}$ is an inverse for $\mathbf{A}^{-1}$. Hence $\mathbf{A}^{-1}$ is invertible, which is equivalent to nonsingularity. □

Solution to Ex. 3.5.2. Let $\mathbf{A}$ be positive definite and consider the following: If $\mathbf{A}$ is singular, then there exists nonzero $\mathbf{x}$ with $\mathbf{A}\mathbf{x} = \mathbf{0}$ (see fact 3.1.3 on page 50). But then $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 0$ for nonzero $\mathbf{x}$. Contradiction. □

Solution to Ex. 3.5.3. If $\mathbf{x} = \mathbf{e}_n$, then $\mathbf{x}^\top \mathbf{A}\mathbf{x} = a_{nn}$. The claim follows. □

Solution to Ex. 3.5.4. Let $\mathbf{A}$, $\mathbf{x}$ and $\mathbf{z}$ be as stated in the question. Suppose, to the contrary, that $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$. Then $\mathbf{A}(\mathbf{x} - \mathbf{z}) = \mathbf{0}$, which, by fact 3.1.2 on page 50, implies $\mathbf{x} - \mathbf{z} = \mathbf{0}$, or $\mathbf{x} = \mathbf{z}$. Contradiction. □

Solution to Ex. 3.5.5. Let $\mathbf{A}$ be a nonsingular matrix and let T be the linear map associated with $\mathbf{A}$ via $T\mathbf{x} = \mathbf{A}\mathbf{x}$. Since A is nonsingular, T is also, by definition, nonsingular and hence, by fact 2.1.9 on page 26, has a nonsingular inverse T^{-1}. Being nonsingular, T^{-1} is necessarily linear, and hence, by theorem 3.1.1 on page 48, there exists a matrix $\mathbf{B}$ such that $T^{-1}\mathbf{x} = \mathbf{B}\mathbf{x}$ for all $\mathbf{x}$. By the definition of the inverse, we

6. Hint: Look at fact 3.3.9 and the argument that follows it.

have $\mathbf{ABx} = T(T^{-1}(\mathbf{x})) = \mathbf{x} = \mathbf{Ix}$. Since this holds for any $\mathbf{x}$, we have $\mathbf{AB} = \mathbf{I}$ (see ex. 3.5.13). A similar argument shows that $\mathbf{BA} = \mathbf{I}$.

Regarding the second claim, $\mathbf{A}^{-1}\mathbf{b}$ is a solution to $\mathbf{Ax} = \mathbf{b}$, since $\mathbf{AA}^{-1}\mathbf{b} = \mathbf{Ib} = \mathbf{b}$. Uniqueness follows from fact 3.1.3 (and, in particular, the fact that nonsingularity of $\mathbf{A}$ includes the implication that the map $\mathbf{x} \mapsto \mathbf{Ax}$ is one-to-one). □

Solution to Ex. 3.5.8. Since the columns of $\mathbf{A}$ consist of K vectors in $\mathbb{R}^N$, the fact that $K > N$ implies that not all of the columns of $\mathbf{A}$ are linearly independent. (Recall theorem 2.1.3 on page 20.) It follows that $\mathbf{Az} = \mathbf{0}$ for some nonzero $\mathbf{z}$ in $\mathbb{R}^K$, and hence $\mathbf{A}\lambda\mathbf{z} = \mathbf{0}$ for any scalar λ. Now suppose that some $\mathbf{x}$ solves $\mathbf{Ax} = \mathbf{b}$. Then, for any $\lambda \in \mathbb{R}$, we have $\mathbf{Ax} + \mathbf{A}\lambda\mathbf{z} = \mathbf{A}(\mathbf{x} + \lambda\mathbf{z}) = \mathbf{b}$. This proves the claim. □

Solution to Ex. 3.5.14. The solutions are as follows: (1) $\mathbf{I}$ is full column rank because its columns are the canonical basis vectors, which are independent. (2) By definition, $\mathbf{B}$ is the inverse of $\mathbf{A}$ if $\mathbf{BA} = \mathbf{AB} = \mathbf{I}$. It follows immediately that $\mathbf{I}$ is the inverse of itself. (3) A sufficient condition is $\alpha > 0$. If this holds, then given $\mathbf{x} \neq \mathbf{0}$, we have $\mathbf{x}^\top \alpha \mathbf{I}\mathbf{x} = \alpha\|\mathbf{x}\|^2 > 0$. □

Solution to Ex. 3.5.15. First, $\mathbf{X}$ is symmetric because

$$\mathbf{X}^\top = (\mathbf{I} - 2\mathbf{uu}^\top)^\top = \mathbf{I} - 2(\mathbf{uu}^\top)^\top = \mathbf{I} - 2(\mathbf{u}^\top)^\top\mathbf{u}^\top = \mathbf{I} - 2\mathbf{uu}^\top = \mathbf{X}$$

Second, $\mathbf{XX} = \mathbf{I}$ because

$$\begin{aligned}\mathbf{XX} = (\mathbf{I} - 2\mathbf{uu}^\top)(\mathbf{I}^\top - 2\mathbf{uu}^\top) &= \mathbf{II} - 2\mathbf{I}2\mathbf{uu}^\top + (2\mathbf{uu}^\top)(2\mathbf{uu}^\top) \\ &= \mathbf{I} - 4\mathbf{uu}^\top + 4\mathbf{uu}^\top\mathbf{uu}^\top = \mathbf{I} - 4\mathbf{uu}^\top + 4\mathbf{uu}^\top = \mathbf{I}\end{aligned}$$

The second last equality is due to the assumption that $\mathbf{u}^\top\mathbf{u} = \|\mathbf{u}\|^2 = 1$. □

Solution to Ex. 3.5.17. Let $\mathbf{A}$ be $N \times N$ and let $\mathbf{I}$ be the $N \times N$ identity. We have

$$\begin{aligned}\det(\mathbf{A} - \lambda\mathbf{I}) = 0 &\iff \mathbf{A} - \lambda\mathbf{I} \text{ is singular} \\ &\iff \exists\, \mathbf{x} \neq \mathbf{0} \text{ s.t. } (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \iff \exists\, \mathbf{x} \neq \mathbf{0} \text{ s.t. } \mathbf{Ax} = \lambda\mathbf{x}\end{aligned}$$

In other words, λ is an eigenvalue of $\mathbf{A}$. □

Solution to Ex. 3.5.19. Suppose that $\mathbf{A}$ is both idempotent and nonsingular. From idempotence we have $\mathbf{AA} = \mathbf{A}$. Premultiplying by $\mathbf{A}^{-1}$ gives $\mathbf{A} = \mathbf{I}$. □

Solution to Ex. 3.5.21. The claim is true because

$$\begin{aligned}
(\mathbf{A} + \mathbf{U}\mathbf{W}) & \left[\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U} \left(\mathbf{I} + \mathbf{W}\mathbf{A}^{-1}\mathbf{U} \right)^{-1} \mathbf{W}\mathbf{A}^{-1} \right] \\
& = \mathbf{I} + \mathbf{U}\mathbf{W}\mathbf{A}^{-1} - (\mathbf{U} + \mathbf{U}\mathbf{W}\mathbf{A}^{-1}\mathbf{U})(\mathbf{I} + \mathbf{W}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{W}\mathbf{A}^{-1} \\
& = \mathbf{I} + \mathbf{U}\mathbf{W}\mathbf{A}^{-1} - \mathbf{U}(\mathbf{I} + \mathbf{W}\mathbf{A}^{-1}\mathbf{U})(\mathbf{I} + \mathbf{W}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{W}\mathbf{A}^{-1} \\
& = \mathbf{I} + \mathbf{U}\mathbf{W}\mathbf{A}^{-1} - \mathbf{U}\mathbf{W}\mathbf{A}^{-1} = \mathbf{I}
\end{aligned}$$

□

Solution to Ex. 3.5.22. Let $\mathbf{Q}$ be an orthogonal matrix with columns $\mathbf{u}_1, \ldots, \mathbf{u}_N$. By the definition of matrix multiplication, the m, nth element of $\mathbf{Q}^\top \mathbf{Q}$ is $\mathbf{u}_m^\top \mathbf{u}_n$, which is 1 if $m = n$ and zero otherwise. Hence $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$. It follows from fact 3.1.4 on page 51 that $\mathbf{Q}^\top$ is the inverse of $\mathbf{Q}$.

To see that $\det \mathbf{Q} \in \{-1, 1\}$, apply the results of fact 3.1.6 (page 52) and fact 3.2.5 (page 55) to the equality $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$ to obtain $\det(\mathbf{Q})^2 = 1$. The claim follows. □

Solution to Ex. 3.5.23. Suppose that $\mathbf{A}$ is symmetric with eigenvalues $\lambda_1, \ldots, \lambda_N$. By theorem 3.3.4 we can decompose it as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, where $\mathbf{\Lambda}$ is the diagonal matrix formed from eigenvalues and $\mathbf{Q}$ is an orthogonal matrix. Fixing $\mathbf{x} \in \mathbb{R}^N$ and letting $\mathbf{y} := \mathbf{Q}^\top \mathbf{x}$, we have

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = (\mathbf{Q}^\top \mathbf{x})^\top \mathbf{\Lambda} (\mathbf{Q}^\top \mathbf{x}) = \mathbf{y}^\top \mathbf{\Lambda} \mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_N y_N^2 \tag{3.19}$$

Suppose that all eigenvalues are positive. Take $\mathbf{x}$ to be nonzero. The vector $\mathbf{y}$ must be nonzero (why?), and it follows from (3.19) that $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$. Hence $\mathbf{A}$ is positive definite as claimed.

Conversely, suppose that $\mathbf{A}$ is positive definite. Fix $n \leqslant N$ and set $\mathbf{x} = \mathbf{Q}\mathbf{e}_n$. Evidently $\mathbf{x}$ is nonzero (why?). Hence $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$. Since $\mathbf{Q}^\top$ is the inverse of $\mathbf{Q}$, it follows that

$$\lambda_n = \mathbf{e}_n^\top \mathbf{\Lambda} \mathbf{e}_n = (\mathbf{Q}^\top \mathbf{x})^\top \mathbf{\Lambda} \mathbf{Q}^\top \mathbf{x} = \mathbf{x}^\top \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top \mathbf{x} = \mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$$

Since n was arbitrary, all eigenvalues are positive. □

Solution to Ex. 3.5.24. Fixing $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and letting $\mathbf{z} := \mathbf{x} - \mathbf{y}$ we have

$$\|\mathbf{Q}\mathbf{x} - \mathbf{Q}\mathbf{y}\|^2 = \|\mathbf{Q}\mathbf{z}\|^2 = (\mathbf{Q}\mathbf{z})^\top \mathbf{Q}\mathbf{z} = \mathbf{z}^\top \mathbf{Q}^\top \mathbf{Q} \mathbf{z} = \mathbf{z}^\top \mathbf{z} = \|\mathbf{z}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2$$

□

Solution to Ex. 3.5.26. If $N = K$, then, in view of the full column rank assumption and theorem 3.1.2 on page 51, the matrix $\mathbf{A}$ is nonsingular. By fact 3.2.5 on page 55, $\mathbf{A}^\top$ is likewise nonsingular. Applying the usual rule for inverse of products (fact 3.1.5

on page 52), we have

$$\hat{\mathbf{x}} = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{b} = \mathbf{A}^{-1}(\mathbf{A}^\top)^{-1}\mathbf{A}^\top \mathbf{b} = \mathbf{A}^{-1}\mathbf{b}$$

This is the solution to the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ when $\mathbf{A}$ is square and invertible. □

Solution to Ex. 3.5.27. Let S and $\mathbf{P}$ be as in the statement of the exercise and let $K := \dim S$. We aim to show that trace $\mathbf{P} = K$. Let $\mathbf{B}$ be a matrix such that its columns form a basis for S. By the definition of dimension, $\mathbf{B}$ has K columns. Applying (3.8), we have

$$\operatorname{trace} \mathbf{P} = \operatorname{trace}(\mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top)$$

Recalling fact 3.2.3, we rearrange to get

$$\operatorname{trace}[\mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top] = \operatorname{trace}[(\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top \mathbf{B}] = \operatorname{trace} \mathbf{I}$$

where $\mathbf{I}$ is the $K \times K$ identity. The claim follows. □

Solution to Ex. 3.5.28. Let $\mathbf{A} = \mathbf{B}^\top \mathbf{B}$. It suffices to show that $\mathbf{A}$ is positive definite, since this implies that its determinant is strictly positive, and any matrix with nonzero determinant is nonsingular. To see that $\mathbf{A}$ is positive definite, pick any $\mathbf{b} \neq \mathbf{0}$. We must show that $\mathbf{b}^\top \mathbf{A}\mathbf{b} > 0$. To see this, observe that

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

By the properties of norms, the last term is zero only when $\mathbf{B}\mathbf{b} = \mathbf{0}$. But this is not true because $\mathbf{b} \neq \mathbf{0}$ and $\mathbf{B}$ is full column rank (see theorem 2.1.1). □

Solution to Ex. 3.5.30. Let S and $\mathbf{P}$ be as stated in theorem 3.3.1 and let $\mathbf{B}$ be a matrix such that the columns of $\mathbf{B}$ form a basis of S. Fix $\mathbf{y} \in \mathbb{R}^N$. The claim is that $\hat{\mathbf{y}} := \mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top \mathbf{y}$ is the orthogonal projection of $\mathbf{y}$ onto S. To verify this, we need to show that

(i) $\hat{\mathbf{y}} \in S$ and

(ii) $\mathbf{y} - \hat{\mathbf{y}} \perp S$.

Part (i) is true because $\hat{\mathbf{y}}$ can be written as $\hat{\mathbf{y}} = \mathbf{B}\mathbf{x}$ where $\mathbf{x} := (\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top \mathbf{y}$. The vector $\mathbf{B}\mathbf{x}$ is a linear combination of the columns of $\mathbf{B}$. Since these columns form a basis of S, they must lie in S. Hence $\hat{\mathbf{y}} \in S$ as claimed.

Regarding (ii), from the assumption that $\mathbf{B}$ gives a basis for S, all points in S have the form $\mathbf{B}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^K$. Thus (ii) translates to the claim that

$$\mathbf{y} - \mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top \mathbf{y} \perp \mathbf{B}\mathbf{x} \quad \text{for all } \mathbf{x} \in \mathbb{R}^K$$

This is true because if $\mathbf{x} \in \mathbb{R}^K$, then

$$(\mathbf{B}\mathbf{x})^\top[\mathbf{y} - \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{y}] = \mathbf{x}^\top[\mathbf{B}^\top\mathbf{y} - \mathbf{B}^\top\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{y}] = \mathbf{x}^\top[\mathbf{B}^\top\mathbf{y} - \mathbf{B}^\top\mathbf{y}] = 0 \quad \square$$

Solution to Ex. 3.5.31. These are straightforward. For example,

$$\mathbf{M}\mathbf{B} = \mathbf{B} - \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{B} = \mathbf{0} \quad \square$$

Solution to Ex. 3.5.32. Let $\mathbf{A}$, $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{b} \in \mathbb{R}^N$ be as in the statement of the exercise. The claim is that $\hat{\mathbf{x}}$ defined in (3.12) is equal to $\tilde{\mathbf{x}} := \mathbf{R}^{-1}\mathbf{Q}^\top\mathbf{b}$. To show this, in view of linear independence of the columns of $\mathbf{A}$, it suffices to show that $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}\hat{\mathbf{x}}$, or

$$\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b} = \mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{Q}^\top\mathbf{b}$$

After simplifying, we see it suffices to show that $\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top = \mathbf{Q}\mathbf{Q}^\top$. Since $\mathbf{A}$ and $\mathbf{Q}$ have the same column space, this follows from theorem 3.3.1 on page 60. $\square$

Solution to Ex. 3.5.33. Regarding (i), the expression for $\mathbf{P}\mathbf{y}$ given in exercise 3.5.33 can also be written as $\mathbf{x}(\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top\mathbf{y}$. Since $\mathbf{x}$ is a basis for span$\{\mathbf{x}\}$, the validity of this expression as the projection onto span$\{\mathbf{x}\}$ follows immediately from theorem 3.3.1. Regarding (ii), recall that orthogonal projections contract norms, so that, in particular, $\|\mathbf{P}\mathbf{y}\| \leqslant \|\mathbf{y}\|$ must hold. Using our expression for $\mathbf{P}\mathbf{y}$ from (i) and rearranging gives the desired bound $|\mathbf{x}^\top\mathbf{y}| \leqslant \|\mathbf{x}\|\|\mathbf{y}\|$. $\square$

Solution to Ex. 3.5.34. By the definition of orthogonal projection, $\mathbf{P}\mathbf{b} \in$ colspace $\mathbf{A}$, and hence there exists a vector $\mathbf{x}$ such that $\mathbf{P}\mathbf{b} = \mathbf{A}\mathbf{x}$. Since $\mathbf{A}$ has linearly dependent columns, there exists a nonzero vector $\mathbf{a}$ such that $\mathbf{A}\mathbf{a} = \mathbf{0}$. Hence $\mathbf{A}\lambda\mathbf{a} = \mathbf{0}$ for all $\lambda \in \mathbb{R}$. For each such λ we have $\mathbf{P}\mathbf{b} = \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{A}\lambda\mathbf{a} = \mathbf{A}(\mathbf{x} + \lambda\mathbf{a})$. This proves the claim. $\square$

Solution to Ex. 3.5.35. By the spectral decomposition theorem (page 66), we know that $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is the diagonal matrix formed from the eigenvalues of $\mathbf{A}$. It follows (see page 65) that $\mathbf{A}^2 = \mathbf{Q}\mathbf{\Lambda}^2\mathbf{Q}^\top$ and, since $\mathbf{A}$ is idempotent, that $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{Q}\mathbf{\Lambda}^2\mathbf{Q}^\top$. From this we obtain $\mathbf{\Lambda} = \mathbf{\Lambda}^2$. For diagonal matrices, powers are obtained by taking powers of the diagonal elements, from which we get $\lambda_n = \lambda_n^2$ for any eigenvalue λ_n. Hence $\lambda_n \in \{0, 1\}$ as claimed. $\square$