ELEMENTS OF MATRIX ALGEBRA

CHUNG-MING KUAN

Department of Finance
National Taiwan University

September 09, 2009

E-mail: ckuan@ntu.edu.tw; URL: homepage.ntu.edu.tw/~ckuan.

Contents

1 Vector

An n-dimensional *vector* in $\Re^n$ is a collection of n real numbers. An n-dimensional *row vector* is written as $(u_1, u_2, \ldots, u_n)$, and the corresponding *column vector* is

$$\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

Clearly, a vector reduces to a *scalar* when $n = 1$. A vector can also be interpreted as a point in a system of coordinate axes, such that its components u_i represent the corresponding coordinates. In what follows, we use standard English and Greek alphabets to denote scalars and those alphabets in boldface to denote vectors.

1.1 Vector Operations

Consider vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ in $\Re^n$ and scalars h and k. Two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are said to be *equal* if they are the same componentwise, i.e., $u_i = v_i$, $i = 1, \ldots, n$. Thus,

$$(u_1, u_2, \ldots, u_n) \neq (u_2, u_1, \ldots, u_n),$$

unless $u_1 = u_2$. Also, (u_1, u_2), $(-u_1, -u_2)$ $(-u_1, u_2)$ and $(u_1, -u_2)$ are not equal. In fact, they are four vectors pointing to different directions. This shows that direction is crucial in determining a vector. This is in contrast with the quantities such as area and length for which direction is of no relevance.

The *sum* of $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as

$$\boldsymbol{u} + \boldsymbol{v} = (u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n),$$

and the *scalar multiple* of $\boldsymbol{u}$ is

$$h\boldsymbol{u} = (hu_1, hu_2, \ldots, hu_n).$$

Moreover,

1. $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$;

2. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$;

3. $h(k\boldsymbol{u}) = (hk)\boldsymbol{u}$;

4. $h(\boldsymbol{u} + \boldsymbol{v}) = h\boldsymbol{u} + h\boldsymbol{v}$;

5. $(h + k)\boldsymbol{u} = h\boldsymbol{u} + k\boldsymbol{u}$;

The *zero vector* $\boldsymbol{o}$ is the vector with all elements zero, so that for any vector $\boldsymbol{u}$, $\boldsymbol{u}+\boldsymbol{o} = \boldsymbol{u}$. As $\boldsymbol{u} + (-\boldsymbol{u}) = \boldsymbol{o}$, $-\boldsymbol{u}$ is also known as the *negative* (*additive inverse*) of $\boldsymbol{u}$. Note that the negative of $\boldsymbol{u}$ is a vector pointing to the opposite direction of $\boldsymbol{u}$.

1.2 Euclidean Inner Product and Norm

The *Euclidean inner product* of $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n.$$

Euclidean inner products have the following properties:

1. $\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{v} \cdot \boldsymbol{u}$;
2. $(\boldsymbol{u} + \boldsymbol{v}) \cdot \boldsymbol{w} = \boldsymbol{u} \cdot \boldsymbol{w} + \boldsymbol{v} \cdot \boldsymbol{w}$;
3. $(h\boldsymbol{u}) \cdot \boldsymbol{v} = h(\boldsymbol{u} \cdot \boldsymbol{v})$, where h is a scalar;
4. $\boldsymbol{u} \cdot \boldsymbol{u} \geq 0$; $\boldsymbol{u} \cdot \boldsymbol{u} = 0$ if, and only if, $\boldsymbol{u} = \boldsymbol{o}$.

The *norm* of a vector is a non-negative real number characterizing its magnitude. A commonly used norm is the *Euclidean norm*:

$$\|\boldsymbol{u}\| = (u_1^2 + \cdots + u_n^2)^{1/2} = (\boldsymbol{u} \cdot \boldsymbol{u})^{1/2}.$$

Taking $\boldsymbol{u}$ as a point in the standard Euclidean coordinate system, the Euclidean norm of $\boldsymbol{u}$ is just the familiar Euclidean distance between this point and the origin. There are other norms; for example, the *maximum norm* of a vector is defined as the largest absolute value of its components:

$$\|\boldsymbol{u}\|_\infty = \max(|u_1|, |u_2|, \cdots, |u_n|).$$

Note that for any norm $\|\cdot\|$ and a scalar h, $\|h\boldsymbol{u}\| = |h| \, \|\boldsymbol{u}\|$. A norm may also be viewed as a generalization of the usual notion of "length;" different norms are just different ways to describe the "length."

1.3 Unit Vector

A vector is said to be a *unit vector* if its norm equals one. Two unit vectors are said to be *orthogonal* if their inner product is zero (see also Section 1.4). For example, $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, and $(0.267, 0.534, 0.802)$ are all unit vectors in $\Re^3$, but only the first three vectors are mutually orthogonal. Orthogonal unit vectors are also known as *orthonormal vectors*. In particular, $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ are orthonormal and referred to as *Cartesian unit vectors*. It is also easy to see that any non-zero vector can be *normalized* to unit length. To see this, observe that for any $\boldsymbol{u} \neq \boldsymbol{o}$,

$$\left\| \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \right\| = \frac{1}{\|\boldsymbol{u}\|} \|\boldsymbol{u}\| = 1.$$

That is, any vector $\boldsymbol{u}$ divided by its norm (i.e., $\boldsymbol{u}/\|\boldsymbol{u}\|$) has norm one.

Any n-dimensional vector can be represented as a linear combination of n orthonormal vectors:

$$\begin{aligned}
\boldsymbol{u} &= (u_1, u_2, \dots, u_n) \\
&= (u_1, 0, \dots, 0) + (0, u_2, 0, \dots, 0) + \cdots + (0, \dots, 0, u_n) \\
&= u_1(1, 0, \dots, 0) + u_2(0, 1, \dots, 0) + \cdots + u_n(0, \dots, 0, 1).
\end{aligned}$$

Hence, orthonormal vectors can be viewed as orthogonal coordinate axes of unit length. We could, of course, change the coordinate system without affecting the vector. For example, $\boldsymbol{u} = (1, 1/2)$ is a vector in the Cartesian coordinate system:

$$\boldsymbol{u} = (1, 1/2) = 1(1, 0) + \frac{1}{2}(0, 1),$$

and it can also be expressed in terms of the orthogonal vectors $(2, 0)$ and $(0, 3)$ as

$$(1, 1/2) = \frac{1}{2}(2, 0) + \frac{1}{6}(0, 3).$$

Thus, $\boldsymbol{u} = (1, 1/2)$ is also the vector $(1/2, 1/6)$ in the coordinate system of the vectors $(2, 0)$ and $(0, 3)$. As $(2, 0)$ and $(0, 3)$ can be expressed in terms of Cartesian unit vectors, it is typical to consider only the Cartesian coordinate system.

1.4 Direction Cosine

Given $\boldsymbol{u}$ in $\Re^n$, let θ_i denote the angle between $\boldsymbol{u}$ and the i^{th} axis. The *direction cosines* of $\boldsymbol{u}$ are

$$\cos\theta_i := u_i/\|\boldsymbol{u}\|, \qquad i = 1, \dots, n.$$

Clearly,

$$\sum_{i=1}^{n} \cos^2\theta_i = \sum_{i=1}^{n} u_i^2/\|\boldsymbol{u}\|^2 = 1.$$

Note that for any scalar non-zero h, $h\boldsymbol{u}$ has the direction cosines:

$$\cos\theta_i = hu_i/\|h\boldsymbol{u}\| = \pm\left(u_i/\|\boldsymbol{u}\|\right), \qquad i = 1, \dots, n.$$

That is, direction cosines are independent of vector magnitude; only the sign of h (direction) matters.

Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be two vectors in $\Re^n$ with direction cosines r_i and s_i, $i = 1, \ldots, n$, and let θ denote the angle between $\boldsymbol{u}$ and $\boldsymbol{v}$. Note that $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{u} - \boldsymbol{v}$ form a triangle. Then by the law of cosine,

$$\|\boldsymbol{u} - \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 - 2\|\boldsymbol{u}\|\,\|\boldsymbol{v}\| \cos\theta,$$

or equivalently,

$$\cos\theta = \frac{\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 - \|\boldsymbol{u} - \boldsymbol{v}\|^2}{2\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|}.$$

The numerator above can be expressed as

$$\begin{aligned}
&\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 - \|\boldsymbol{u} - \boldsymbol{v}\|^2 \\
&\qquad = \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 - \|\boldsymbol{u}\|^2 - \|\boldsymbol{v}\|^2 + 2\boldsymbol{u} \cdot \boldsymbol{v} \\
&\qquad = 2\boldsymbol{u} \cdot \boldsymbol{v}.
\end{aligned}$$

Hence,

$$\cos\theta = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|}.$$

We have proved:

Theorem 1.1 *For two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in $\Re^n$,*

$$\boldsymbol{u} \cdot \boldsymbol{v} = \|\boldsymbol{u}\|\,\|\boldsymbol{v}\| \cos\theta,$$

where θ is the angle between $\boldsymbol{u}$ and $\boldsymbol{v}$.

When $\theta = 0$ (π), $\boldsymbol{u}$ and $\boldsymbol{v}$ are on the same "line" with the same (opposite) direction. In this case, $\boldsymbol{u}$ and $\boldsymbol{v}$ are said to be *linearly dependent* (*collinear*), and $\boldsymbol{u}$ can be written as $h\boldsymbol{v}$ for some scalar h. When $\theta = \pi/2$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are said to be *orthogonal*. As $\cos(\pi/2) = 0$, two non-zero vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal if, and only if, $\boldsymbol{u} \cdot \boldsymbol{v} = 0$.

As $-1 \leq \cos\theta \leq 1$, we immediately have from Theorem 1.1 that:

Theorem 1.2 (Cauchy-Schwartz Inequality) *Given two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$,*

$$|\boldsymbol{u} \cdot \boldsymbol{v}| \leq \|\boldsymbol{u}\|\|\boldsymbol{v}\|;$$

the equality holds when $\boldsymbol{u}$ and $\boldsymbol{v}$ are linearly dependent.

By the Cauchy-Schwartz inequality,

$$\begin{aligned}\|\boldsymbol{u}+\boldsymbol{v}\|^2 &= (\boldsymbol{u}\cdot\boldsymbol{u})+(\boldsymbol{v}\cdot\boldsymbol{v})+2(\boldsymbol{u}\cdot\boldsymbol{v})\\ &\le \|\boldsymbol{u}\|^2+\|\boldsymbol{v}\|^2+2\|\boldsymbol{u}\|\|\boldsymbol{v}\|\\ &= (\|\boldsymbol{u}\|+\|\boldsymbol{v}\|)^2.\end{aligned}$$

This establishes the well known triangle inequality.

Theorem 1.3 (Triangle Inequality) *Given two vectors* $\boldsymbol{u}$ *and* $\boldsymbol{v}$*,*

$$\|\boldsymbol{u}+\boldsymbol{v}\| \le \|\boldsymbol{u}\|+\|\boldsymbol{v}\|;$$

the equality holds when $\boldsymbol{u}=h\boldsymbol{v}$ *and* $h>0$*.*

If $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal, we have

$$\|\boldsymbol{u}+\boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2+\|\boldsymbol{v}\|^2,$$

the *generalized Pythagoras theorem.*

1.5 Statistical Applications

Given a random variable X with n observations $x_1,\ldots,x_n$, different *statistics* can be used to summarize the information contained in this sample. An important statistic is the *sample average* of x_i which labels the "location" of these observations. Let $\boldsymbol{x}$ denote the vector of these n observations. The sample average of x_i is

$$\bar{\boldsymbol{x}} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(\boldsymbol{x}\cdot\boldsymbol{\ell}),$$

where $\boldsymbol{\ell}$ is the vector of ones. Another important statistic is the *sample variance* of x_i which describes the "dispersion" of these observations. Let $\boldsymbol{x}^* = \boldsymbol{x}-\bar{\boldsymbol{x}}\boldsymbol{\ell}$ be the vector of deviations from the sample average. The sample variance of x_i is

$$s_x^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}})^2 = \frac{1}{n-1}\|\boldsymbol{x}^*\|^2.$$

Note that s_x^2 is *invariant* with respect to scalar addition, in the sense that $s_x^2 = s_{x+a}^2$ for any scalar a. By contrast, the *sample second moment* of $\boldsymbol{x}$,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\|\boldsymbol{x}\|^2,$$

is not *invariant* with respect to scalar addition. Also note that sample variance is divided by $n-1$ rather than n. This is because

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}}) = 0,$$

so that any component of $\boldsymbol{x}^*$ depends on the remaining $n-1$ components. The square root of s_x^2 is called the *standard deviation* of x_i, denoted as s_x.

For two random variables X and Y with the vectors of observations $\boldsymbol{x}$ and $\boldsymbol{y}$, their *sample covariance* characterizes the co-variation of these observations:

$$s_{x,y} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}) = \frac{1}{n-1}(\boldsymbol{x}^* \cdot \boldsymbol{y}^*),$$

where $\boldsymbol{y}^* = \boldsymbol{y} - \bar{\boldsymbol{y}}\boldsymbol{\ell}$. This statistic is again invariant with respect to scalar addition.

Both sample variance and covariance are not invariant with respect to constant multiplication (i.e., not scale invariant). The sample covariance normalized by corresponding standard deviations is called the *sample correlation coefficient*:

$$\begin{aligned} r_{x,y} &:= \frac{\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}})}{[\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})^2]^{1/2}[\sum_{i=1}^{n}(y_i - \bar{\boldsymbol{y}})^2]^{1/2}} \\ &= \frac{\boldsymbol{x}^* \cdot \boldsymbol{y}^*}{\|\boldsymbol{x}^*\|\|\boldsymbol{y}^*\|} \\ &= \cos\theta^*, \end{aligned}$$

where θ^* is the angle between $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$. Clearly, $r_{x,y}$ is scale invariant and bounded between -1 and 1.

Exercises

1.1 Let $\boldsymbol{u} = (1, -3, 2)$ and $\boldsymbol{v} = (4, 2, 1)$. Draw a figure to show $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{u}+\boldsymbol{v}$, and $\boldsymbol{u}-\boldsymbol{v}$.

1.2 Find the norms and pairwise inner products of the vectors: $(-3, 2, 4)$, $(-3, 0, 0)$, and $(0, 0, 4)$. Also normalize them to unit vectors.

1.3 Find the angle between the vectors $(1, 2, 0, 3)$ and $(2, 4, -1, 1)$.

1.4 Find two unit vectors that are orthogonal to $(3, -2)$.

1.5 Prove that the zero vector is the only vector with norm zero.

1.6 Given two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, prove that

$$\Big| \|\boldsymbol{u}\| - \|\boldsymbol{v}\| \Big| \le \|\boldsymbol{u} - \boldsymbol{v}\|.$$

Under what condition does this inequality hold as an equality?

1.7 (Minkowski Inequality) Given k vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$, prove that

$$\|\boldsymbol{u}_1 + \boldsymbol{u}_2 + \cdots + \boldsymbol{u}_k\| \le \|\boldsymbol{u}_1\| + \|\boldsymbol{u}_2\| + \cdots + \|\boldsymbol{u}_k\|.$$

1.8 Express the Cauchy-Schwartz and triangle inequalities in terms of sample covariances and standard deviations.

2 Vector Space

A set V of vectors in $\Re^n$ is a *vector space* if it is closed under vector addition and scalar multiplication. That is, for any vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in V, $\boldsymbol{u}+\boldsymbol{v}$ and $h\boldsymbol{u}$ are also in V, where h is a scalar. Note that vectors in $\Re^n$ with the standard operations of addition and scalar multiplication must obey the properties mentioned in Section 1.1. For example, $\Re^n$ and $\{\boldsymbol{o}\}$ are vector spaces, but the set of all points (a,b) with $a \geq 0$ and $b \geq 0$ is not a vector space. (Why?) A set S is a *subspace* of a vector space V if $S \subseteq V$ is closed under vector addition and scalar multiplication. For examples, $\{\boldsymbol{o}\}$, $\Re^3$, lines through the origin, and planes through the origin are all subspaces of $\Re^3$; $\{\boldsymbol{o}\}$, $\Re^2$, and lines through the origin are all subspaces of $\Re^2$.

2.1 The Dimension of a Vector Space

The vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ in a vector space V are said to *span* V if every vector in V can be expressed as a linear combination of these vectors. That is, for any $\boldsymbol{v} \in V$,

$$\boldsymbol{v} = a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2 + \cdots + a_k\boldsymbol{u}_k,$$

where $a_1, \ldots, a_k$ are real numbers. We also say that $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$ form a *spanning set* of V. Intuitively, a spanning set of V contains all the information needed to generate V. It is not difficult to see that, given k spanning vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$, all linear combinations of these vectors form a subspace of V, denoted as $\text{span}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$. In fact, this is the smallest subspace containing $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$. For example, Let $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ be non-collinear vectors in $\Re^3$ with initial points at the origin, then all the linear combinations of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ form a plane through the origin, which is a subspace of $\Re^3$.

Let $S = \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k\}$ be a set of non-zero vectors. Then S is said to be a *linearly independent* set if the only solution to the vector equation

$$a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2 + \cdots + a_k\boldsymbol{u}_k = \boldsymbol{o}$$

is $a_1 = a_2 = \cdots = a_k = 0$; if there are other solutions, S is said to be *linearly dependent*. Clearly, any one of k linearly dependent vectors can be written as a linear combination of the remaining $k-1$ vectors. For example, the vectors: $\boldsymbol{u}_1 = (2,-1,0,3)$, $\boldsymbol{u}_2 = (1,2,5,-1)$, and $\boldsymbol{u}_3 = (7,-1,5,8)$ are linearly dependent because $3\boldsymbol{u}_1 + \boldsymbol{u}_2 - \boldsymbol{u}_3 = \boldsymbol{o}$, and the vectors: $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ are clearly linearly independent. Note that for a set of linearly dependent vectors, there must be at least one "redundant" vector, in the sense that the information contained in this vector is also contained in the remaining vectors. It is also easy to show the following result; see Exercise 2.3.

Theorem 2.1 *Let* $S = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\} \subseteq V$.

(a) *If there exists a subset of* S *which contains* $r \leq k$ *linearly dependent vectors, then* S *is also linearly dependent.*

(b) *If* S *is a linearly independent set, then any subset of* $r \leq k$ *vectors is also linearly independent.*

A set of linearly independent vectors in V is a *basis* for V if these vectors span V. While the vectors of a spanning set of V may be linearly dependent, the vectors of a basis must be linearly independent. Intuitively, we may say that a basis is a spanning set without "redundant" information. A nonzero vector space V is *finite dimensional* if its basis contains a finite number of spanning vectors; otherwise, V is *infinite dimensional*. The *dimension* of a finite dimensional vector space V is the number of vectors in a basis for V. Note that $\{\boldsymbol{o}\}$ is a vector space with dimension zero. As examples we note that $\{(1,0),(0,1)\}$ and $\{(-3,7),(5,5)\}$ are two bases for $\Re^2$. If the dimension of a vector space is known, the result below shows that a set of vectors is a basis if it is either a spanning set or a linearly independent set.

Theorem 2.2 *Let* V *be a* k*-dimensional vector space and* $S = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$. *Then* S *is a basis for* V *provided that either* S *spans* V *or* S *is a set of linearly independent vectors.*

Proof: If S spans V but S is not a basis, then the vectors in S are linearly dependent. We thus have a subset of S that spans V but contains $r < k$ linearly independent vectors. It follows that the dimension of V should be r, contradicting the original hypothesis. Conversely, if S is a linearly independent set but not a basis, then S does not span V. Thus, there must exist $r > k$ linearly independent vectors spanning V. This again contradicts the hypothesis that V is k-dimensional. $\square$

If $S = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$ is a set of linearly independent vectors in a k-dimensional vector space V such that $r < k$, then S is not a basis for V. We can find a vector $\boldsymbol{u}_{r+1}$ which is linearly independent of the vectors in S. By enlarging S to $S' = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r, \boldsymbol{u}_{r+1}\}$ and repeating this step $k - r$ times, we obtain a set of k linearly independent vectors. It follows from Theorem 2.2 that this set must be a basis for V. We have proved:

Theorem 2.3 *Let* V *be a* k*-dimensional vector space and* $S = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$, $r \leq k$, *be a set of linearly independent vectors. Then there exist* $k - r$ *vectors* $\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_k$ *which are linearly independent of* S *such that* $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r, \ldots, \boldsymbol{u}_k\}$ *form a basis for* V.

An implication of this result is that any two bases must have the same number of vectors.

2.2 The Sum and Direct Sum of Vector Spaces

Given two vector spaces U and V, the set of all vectors belonging to both U and V is called the *intersection* of U and V, denoted as $U \cap V$, and the set of all vectors $\boldsymbol{u} + \boldsymbol{v}$, where $\boldsymbol{u} \in U$ and $\boldsymbol{v} \in V$, is called the *sum* or *union* of U and V, denoted as $U \cup V$.

Theorem 2.4 *Given two vector spaces U and V, let $\dim(U) = m$, $\dim(V) = n$, $\dim(U \cap V) = k$, and $\dim(U \cup V) = p$, where $\dim(\cdot)$ denotes the dimension of a vector space. Then, $p = m + n - k$.*

Proof: Also let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_k$ denote the basis vectors of $U \cap V$. The basis for U can be written as

$$S(U) = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_k, \boldsymbol{u}_1, \dots, \boldsymbol{u}_{m-k}\},$$

where $\boldsymbol{u}_i$ are some vectors not in $U \cap V$, and the basis for V can be written as

$$S(V) = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_k, \boldsymbol{v}_1, \dots, \boldsymbol{v}_{n-k}\},$$

where $\boldsymbol{v}_i$ are some vectors not in $U \cap V$. It can be seen that the vectors in $S(U)$ and $S(V)$ form a spanning set for $U \cup V$. The assertion follows if these vectors form a basis. We therefore must show that these vectors are linearly independent. Consider an arbitrary linear combination:

$$a_1\boldsymbol{w}_1 + \cdots + a_k\boldsymbol{w}_k + b_1\boldsymbol{u}_1 + \cdots + b_{m-k}\boldsymbol{u}_{m-k} + c_1\boldsymbol{v}_1 + \cdots + c_{n-k}\boldsymbol{v}_{n-k} = 0.$$

Then,

$$a_1\boldsymbol{w}_1 + \cdots + a_k\boldsymbol{w}_k + b_1\boldsymbol{u}_1 + \cdots + b_{m-k}\boldsymbol{u}_{m-k} = -c_1\boldsymbol{v}_1 - \cdots - c_{n-k}\boldsymbol{v}_{n-k},$$

where the left-hand side is a vector in U. It follows that the right-hand side is also in U. Note, however, that a linear combination of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n-k}$ should be a vector in V but not in U. Hence, the only possibility is that the right-hand side is the zero vector. This implies that the coefficients c_i must be all zeros because $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n-k}$ are linearly independent. Consequently, all a_i and b_i must also be all zeros. $\square$

When $U \cap V = \{\boldsymbol{o}\}$, $U \cup V$ is called the *direct sum* of U and V, denoted as $U \oplus V$. It follows from Theorem 2.4 that the dimension of the direct sum of U and V is

$$\dim(U \oplus V) = \dim(U) + \dim(V).$$

If $\boldsymbol{w} \in U \oplus V$, then $\boldsymbol{w} = \boldsymbol{u}_1 + \boldsymbol{v}_1$ for some $\boldsymbol{u}_1 \in U$ and $\boldsymbol{v}_1 \in V$. If one can also write $\boldsymbol{w} = \boldsymbol{u}_2 + \boldsymbol{v}_2$, where $\boldsymbol{u}_1 \neq \boldsymbol{u}_2 \in U$ and $\boldsymbol{v}_1 \neq \boldsymbol{v}_2 \in V$, then $\boldsymbol{u}_1 - \boldsymbol{u}_2 = \boldsymbol{v}_2 - \boldsymbol{v}_1$ is a non-zero vector belonging to both U and V. But this is not possible by the definition of direct sum. This shows that the decomposition of $\boldsymbol{w} \in U \oplus V$ must be unique.

Theorem 2.5 *Any vector $\boldsymbol{w} \in U \oplus V$ can be written uniquely as $\boldsymbol{w} = \boldsymbol{u} + \boldsymbol{v}$, where $\boldsymbol{u} \in U$ and $\boldsymbol{v} \in V$.*

More generally, given vector spaces V_i, $i = 1, \ldots, n$, such that $V_i \cap V_j = \{\boldsymbol{o}\}$ for all $i \neq j$, we have

$$\dim(V_1 \oplus V_2 \oplus \cdots \oplus V_n) = \dim(V_1) + \dim(V_2) + \cdots + \dim(V_n).$$

That is, the dimension of a direct sum is simply the sum of individual dimensions. Theorem 2.5 thus implies that any vector $\boldsymbol{w} \in V_1 \oplus \cdots \oplus V_n$ can be written uniquely as $\boldsymbol{w} = \boldsymbol{v}_1 + \cdots + \boldsymbol{v}_n$, where $\boldsymbol{v}_i \in V_i$, $i = 1, \ldots, n$.

2.3 Orthogonal Basis Vectors

A set of vectors is an *orthogonal set* if all vectors in this set are *mutually orthogonal*; an orthogonal set of unit vectors is an *orthonormal set.* If a vector $\boldsymbol{v}$ is orthogonal to $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$, then $\boldsymbol{v}$ must be orthogonal to any linear combination of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$, and hence the space spanned by these vectors.

A k-dimensional space must contain exactly k mutually orthogonal vectors. Given a k-dimensional space V, consider now an arbitrary linear combination of k mutually orthogonal vectors in V:

$$a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2 + \cdots + a_k\boldsymbol{u}_k = \boldsymbol{o}.$$

Taking inner products with $\boldsymbol{u}_i$, $i = 1, \ldots, k$, we obtain

$$a_1\|\boldsymbol{u}_1\|^2 = \cdots = a_k\|\boldsymbol{u}_k\|^2 = 0.$$

This implies that $a_1 = \cdots = a_k = 0$. Thus, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ are linearly independent and form a basis. Conversely, given a set of k linearly independent vectors, it is always possible to construct a set of k mutually orthogonal vectors; see Section 2.4. These results suggest that we can always consider an orthogonal (or orthonormal) basis; see also Section 1.3.

Two vector spaces U and V are said to be orthogonal if $\boldsymbol{u} \cdot \boldsymbol{v} = 0$ for every $\boldsymbol{u} \in U$ and $\boldsymbol{v} \in V$. For a subspace S of V, its orthogonal complement is a vector space defined as:

$$S^{\perp} := \{\boldsymbol{v} \in V : \boldsymbol{v} \cdot \boldsymbol{s} = 0 \text{ for every } \boldsymbol{s} \in S\} .$$

Thus, S and $S^{\perp}$ are orthogonal, and $S \cap S^{\perp} = \{\boldsymbol{o}\}$ so that $S \cup S^{\perp} = S \oplus S^{\perp}$. Clearly, $S \oplus S^{\perp}$ is a subspace of V. Suppose that V is n-dimensional and S is r-dimensional with $r < n$. Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ be the orthonormal basis of V with $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$ being the orthonormal basis of S. If $\boldsymbol{v} \in S^{\perp}$, it can be written as

$$\boldsymbol{v} = a_1 \boldsymbol{u}_1 + \cdots + a_r \boldsymbol{u}_r + a_{r+1} \boldsymbol{u}_{r+1} + \cdots + a_n \boldsymbol{u}_n.$$

As $\boldsymbol{v} \cdot \boldsymbol{u}_i = a_i$, we have $a_1 = \cdots = a_r = 0$, but a_i need not be 0 for $i = r+1, \ldots, n$. Hence, any vector in $S^{\perp}$ can be expressed as a linear combination of orthonormal vectors $\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_n$. It follows that $S^{\perp}$ is $n - r$ dimensional and

$$\dim(S \oplus S^{\perp}) = \dim(S) + \dim(S^{\perp}) = r + (n - r) = n.$$

That is, $\dim(S \oplus S^{\perp}) = \dim(V)$. This proves the following important result.

Theorem 2.6 *Let V be a vector space and S its subspace. Then, $V = S \oplus S^{\perp}$.*

The corollary below follows from Theorem 2.5 and 2.6.

Corollary 2.7 *Given the vector space V, $\boldsymbol{v} \in V$ can be uniquely expressed as $\boldsymbol{v} = \boldsymbol{s} + \boldsymbol{e}$, where $\boldsymbol{s}$ is in a subspace S and $\boldsymbol{e}$ is in $S^{\perp}$.*

2.4 Orthogonal Projection

Given two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, let $\boldsymbol{u} = \boldsymbol{s} + \boldsymbol{e}$. It turns out that $\boldsymbol{s}$ can be chosen as a scalar multiple of $\boldsymbol{v}$ and $\boldsymbol{e}$ is orthogonal to $\boldsymbol{v}$. To see this, suppose that $\boldsymbol{s} = h\boldsymbol{v}$ and $\boldsymbol{e}$ is orthogonal to $\boldsymbol{v}$. Then

$$\boldsymbol{u} \cdot \boldsymbol{v} = (\boldsymbol{s} + \boldsymbol{e}) \cdot \boldsymbol{v} = (h\boldsymbol{v} + \boldsymbol{e}) \cdot \boldsymbol{v} = h(\boldsymbol{v} \cdot \boldsymbol{v}).$$

This equality is satisfied for $h = (\boldsymbol{u} \cdot \boldsymbol{v})/(\boldsymbol{v} \cdot \boldsymbol{v})$. We have

$$\boldsymbol{s} = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\boldsymbol{v} \cdot \boldsymbol{v}} \boldsymbol{v}, \qquad \boldsymbol{e} = \boldsymbol{u} - \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\boldsymbol{v} \cdot \boldsymbol{v}} \boldsymbol{v}.$$

That is, $\boldsymbol{u}$ can always be decomposed into two orthogonal components $\boldsymbol{s}$ and $\boldsymbol{e}$, where $\boldsymbol{s}$ is known as the *orthogonal projection* of $\boldsymbol{u}$ on $\boldsymbol{v}$ (or the space spanned by $\boldsymbol{v}$) and $\boldsymbol{e}$ is orthogonal to $\boldsymbol{v}$. For example, consider $\boldsymbol{u} = (a, b)$ and $\boldsymbol{v} = (1, 0)$. Then $\boldsymbol{u} = (a, 0)+(0, b)$, where $(a, 0)$ is the orthogonal projection of $\boldsymbol{u}$ on $(1, 0)$ and $(0, b)$ is orthogonal to $(1, 0)$.

More generally, Corollary 2.7 shows that a vector $\boldsymbol{u}$ in an n-dimensional space V can be uniquely decomposed into two orthogonal components: the orthogonal projection of $\boldsymbol{u}$ onto a r-dimensional subspace S and the component in its orthogonal complement. If $S = \mathrm{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$, we can write the orthogonal projection onto S as

$$\boldsymbol{s} = a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_r\boldsymbol{v}_r$$

and $\boldsymbol{e} \cdot \boldsymbol{s} = \boldsymbol{e} \cdot \boldsymbol{v}_i = 0$ for $i = 1, \ldots, r$. When $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$ is an orthogonal set, it can be verified that $a_i = (\boldsymbol{u} \cdot \boldsymbol{v}_i)/(\boldsymbol{v}_i \cdot \boldsymbol{v}_i)$ so that

$$\boldsymbol{s} = \frac{\boldsymbol{u} \cdot \boldsymbol{v}_1}{\boldsymbol{v}_1 \cdot \boldsymbol{v}_1}\,\boldsymbol{v}_1 + \frac{\boldsymbol{u} \cdot \boldsymbol{v}_2}{\boldsymbol{v}_2 \cdot \boldsymbol{v}_2}\,\boldsymbol{v}_2 + \cdots + \frac{\boldsymbol{u} \cdot \boldsymbol{v}_r}{\boldsymbol{v}_r \cdot \boldsymbol{v}_r}\,\boldsymbol{v}_r,$$

and $\boldsymbol{e} = \boldsymbol{u} - \boldsymbol{s}$.

This decomposition is useful because the orthogonal projection of a vector $\boldsymbol{u}$ is the "best approximation" of $\boldsymbol{u}$ in the sense that the distance between $\boldsymbol{u}$ and its orthogonal projection onto S is less than the distance between $\boldsymbol{u}$ and any other vector in S.

Theorem 2.8 *Given a vector space V with a subspace S, let $\boldsymbol{s}$ be the orthogonal projection of $\boldsymbol{u} \in V$ onto S. Then*

$$\|\boldsymbol{u} - \boldsymbol{s}\| \leq \|\boldsymbol{u} - \boldsymbol{v}\|,$$

for any $\boldsymbol{v} \in S$.

Proof: We can write $\boldsymbol{u} - \boldsymbol{v} = (\boldsymbol{u} - \boldsymbol{s}) + (\boldsymbol{s} - \boldsymbol{v})$, where $\boldsymbol{s} - \boldsymbol{v}$ is clearly in S, and $\boldsymbol{u} - \boldsymbol{s}$ is orthogonal to S. Thus, by the generalized Pythagoras theorem,

$$\|\boldsymbol{u} - \boldsymbol{v}\|^2 = \|\boldsymbol{u} - \boldsymbol{s}\|^2 + \|\boldsymbol{s} - \boldsymbol{v}\|^2 \geq \|\boldsymbol{u} - \boldsymbol{s}\|^2.$$

The inequality becomes an equality if, and only if, $\boldsymbol{v} = \boldsymbol{s}$. □

As discussed in Section 2.3, a set of linearly independent vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ can be transformed to an orthogonal basis. The Gram-Schmidt orthogonalization procedure does this by sequentially performing orthogonal projection of each $\boldsymbol{u}_i$ on previously orthogonalized vectors. Specifically,

$$\begin{aligned}
\boldsymbol{v}_1 &= \boldsymbol{u}_1,\\
\boldsymbol{v}_2 &= \boldsymbol{u}_2 - \frac{\boldsymbol{u}_2 \cdot \boldsymbol{v}_1}{\boldsymbol{v}_1 \cdot \boldsymbol{v}_1}\,\boldsymbol{v}_1,\\
\boldsymbol{v}_3 &= \boldsymbol{u}_3 - \frac{\boldsymbol{u}_3 \cdot \boldsymbol{v}_2}{\boldsymbol{v}_2 \cdot \boldsymbol{v}_2}\,\boldsymbol{v}_2 - \frac{\boldsymbol{u}_3 \cdot \boldsymbol{v}_1}{\boldsymbol{v}_1 \cdot \boldsymbol{v}_1}\,\boldsymbol{v}_1,\\
&\ \vdots\\
\boldsymbol{v}_k &= \boldsymbol{u}_k - \frac{\boldsymbol{u}_k \cdot \boldsymbol{v}_{k-1}}{\boldsymbol{v}_{k-1} \cdot \boldsymbol{v}_{k-1}}\,\boldsymbol{v}_{k-1} - \frac{\boldsymbol{u}_k \cdot \boldsymbol{v}_{k-2}}{\boldsymbol{v}_{k-2} \cdot \boldsymbol{v}_{k-2}}\,\boldsymbol{v}_{k-2} - \cdots - \frac{\boldsymbol{u}_k \cdot \boldsymbol{v}_1}{\boldsymbol{v}_1 \cdot \boldsymbol{v}_1}\,\boldsymbol{v}_1.
\end{aligned}$$

It can be easily verified that $\boldsymbol{v}_i$, $i = 1, \ldots, k$, are mutually orthogonal.

2.5 Statistical Applications

Consider two variables with n observations: $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$. One way to describe the relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ is to compute (estimate) a straight line, $\hat{\boldsymbol{y}} = \alpha\boldsymbol{\ell} + \beta\boldsymbol{x}$, that "best" fits the data points (y_i, x_i), $i = 1, \ldots, n$. In the light of Theorem 2.8, this objective can be achieved by computing the orthogonal projection of $\boldsymbol{y}$ onto the space spanned by $\boldsymbol{\ell}$ and $\boldsymbol{x}$.

We first write $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} = \alpha\boldsymbol{\ell} + \beta\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{e}$ is orthogonal to $\boldsymbol{\ell}$ and $\boldsymbol{x}$, and hence $\hat{\boldsymbol{y}}$. To find unknown α and β, note that

$$\begin{aligned}
\boldsymbol{y} \cdot \boldsymbol{\ell} &= \alpha(\boldsymbol{\ell} \cdot \boldsymbol{\ell}) + \beta(\boldsymbol{x} \cdot \boldsymbol{\ell}),\\
\boldsymbol{y} \cdot \boldsymbol{x} &= \alpha(\boldsymbol{\ell} \cdot \boldsymbol{x}) + \beta(\boldsymbol{x} \cdot \boldsymbol{x}).
\end{aligned}$$

Equivalently, we obtain the so-called *normal equations*:

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= n\alpha + \beta \sum_{i=1}^{n} x_i,\\
\sum_{i=1}^{n} x_i y_i &= \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2.
\end{aligned}$$

Solving these equations for α and β we obtain

$$\begin{aligned}
\alpha &= \bar{\boldsymbol{y}} - \beta\bar{\boldsymbol{x}},\\
\beta &= \frac{\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}})}{\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})^2}.
\end{aligned}$$

This is known as the *least squares* method. The resulting straight line, $\hat{\boldsymbol{y}} = \alpha\boldsymbol{\ell} + \beta\boldsymbol{x}$, is the least-squares regression line, and $\boldsymbol{e}$ is the vector of regression residuals. It is evident that the regression line so computed has made $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\| = \|\boldsymbol{e}\|$ (hence the sum of squared errors $\|\boldsymbol{e}\|^2$) as small as possible.

Exercises

2.1 Show that every vector space must contain the zero vector.

2.2 Determine which of the following are subspaces of $\Re^3$ and explain why.

(a) All vectors of the form $(a, 0, 0)$.

(b) All vectors of the form (a, b, c), where $b = a + c$.

2.3 Prove Theorem 2.1.

2.4 Let S be a basis for an n-dimensional vector space V. Show that every set in V with more than n vectors must be linearly dependent.

2.5 Which of the following sets of vectors are bases for $\Re^2$?

(a) $(2, 1)$ and $(3, 0)$.

(b) $(3, 9)$ and $(-4, -12)$.

2.6 The subspace of $\Re^3$ spanned by $(4/5, 0, -3/5)$ and $(0, 1, 0)$ is a plane passing through the origin, denoted as S. Decompose $\boldsymbol{u} = (1, 2, 3)$ as $\boldsymbol{u} = \boldsymbol{w} + \boldsymbol{e}$, where $\boldsymbol{w}$ is the orthogonal projection of $\boldsymbol{u}$ onto S and $\boldsymbol{e}$ is orthogonal to S.

2.7 Transform the basis $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4\}$ into an orthogonal set, where $\boldsymbol{u}_1 = (0, 2, 1, 0)$, $\boldsymbol{u}_2 = (1, -1, 0, 0)$, $\boldsymbol{u}_3 = (1, 2, 0, -1)$, and $\boldsymbol{u}_4 = (1, 0, 0, 1)$.

2.8 Find the least squares regression line of $\boldsymbol{y}$ on $\boldsymbol{x}$ and its residual vector. Compare your result with that of Section 2.5.

3 Matrix

A *matrix* $\boldsymbol{A}$ is an $n \times k$ rectangular array

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{bmatrix}$$

with the (i, j)th element a_{ij}. The i th row of $\boldsymbol{A}$ is $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \ldots, a_{ik})$, and the j th column of $\boldsymbol{A}$ is

$$\boldsymbol{a}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix}.$$

When $n = k = 1$, $\boldsymbol{A}$ becomes a *scalar*; when $n = 1$ ($k = 1$), $\boldsymbol{A}$ is simply a row (column) vector. Hence, a matrix can also be viewed as a collection of vectors. In what follows, a matrix is denoted by capital letters in boldface. It is also quite common to treat a column vector as an $n \times 1$ matrix so that matrix operations can be applied directly.

3.1 Basic Matrix Types

We first discuss some general types of matrices. A matrix is a *square matrix* if the number of rows equals the number of columns. A matrix is a zero matrix if its elements are all zeros. A *diagonal matrix* is a square matrix such that $a_{ij} = 0$ for all $i \neq j$ and $a_{ii} \neq 0$ for some i, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}.$$

When $a_{ii} = c$ for all i, it is a *scalar matrix*; when $c = 1$, it is the *identity matrix*, denoted as $\boldsymbol{I}$. A *lower triangular matrix* is a square matrix of the following form:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix};$$

similarly, an *upper triangular matrix* is a square matrix with the non-zero elements located above the main diagonal. A *symmetric matrix* is a square matrix such that $a_{ij} = a_{ji}$ for all i, j.

3.2 Matrix Operations

Let $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ be two $n \times k$ matrices. $\boldsymbol{A}$ and $\boldsymbol{B}$ are equal if $a_{ij} = b_{ij}$ for all i, j. The sum of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the matrix $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{C}$ with $c_{ij} = a_{ij} + b_{ij}$ for all i, j. Now let $\boldsymbol{A}$ be $n \times k$, $\boldsymbol{B}$ be $k \times m$, and h be a scalar. The scalar multiplication of $\boldsymbol{A}$ is the matrix $h\boldsymbol{A} = \boldsymbol{C}$ with $c_{ij} = ha_{ij}$ for all i, j. The product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the $n \times m$ matrix $\boldsymbol{AB} = \boldsymbol{C}$ with $c_{ij} = \sum_{s=1}^{k} a_{is} b_{sj}$ for all i, j. That is, c_{ij} is the inner product of the i th row of $\boldsymbol{A}$ and the j th column of $\boldsymbol{B}$. This shows that matrix multiplication $\boldsymbol{AB}$ is well defined only when the number of columns of $\boldsymbol{A}$ is the same as the number of rows of $\boldsymbol{B}$.

Let $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ denote matrices and h and k denote scalars. Matrix addition and multiplication have the following properties:

1. $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$;
2. $\boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}) = (\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C}$;
3. $\boldsymbol{A} + \boldsymbol{o} = \boldsymbol{A}$, where $\boldsymbol{o}$ is a zero matrix;
4. $\boldsymbol{A} + (-\boldsymbol{A}) = \boldsymbol{o}$;
5. $h\boldsymbol{A} = \boldsymbol{A}h$;
6. $h(k\boldsymbol{A}) = (hk)\boldsymbol{A}$;
7. $h(\boldsymbol{AB}) = (h\boldsymbol{A})\boldsymbol{B} = \boldsymbol{A}(h\boldsymbol{B})$;
8. $h(\boldsymbol{A} + \boldsymbol{B}) = h\boldsymbol{A} + h\boldsymbol{B}$;
9. $(h + k)\boldsymbol{A} = h\boldsymbol{A} + k\boldsymbol{A}$;
10. $\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$;
11. $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$.

Note, however, that matrix multiplication is *not* commutative, i.e., $\boldsymbol{AB} \neq \boldsymbol{BA}$. For example, if $\boldsymbol{A}$ is $n \times k$ and $\boldsymbol{B}$ is $k \times n$, then $\boldsymbol{AB}$ and $\boldsymbol{BA}$ do not even have the same size. Also note that $\boldsymbol{AB} = \boldsymbol{o}$ does *not* imply that $\boldsymbol{A} = \boldsymbol{o}$ or $\boldsymbol{B} = \boldsymbol{o}$, and that $\boldsymbol{AB} = \boldsymbol{AC}$ does *not* imply $\boldsymbol{B} = \boldsymbol{C}$.

Let $\boldsymbol{A}$ be an $n \times k$ matrix. The *transpose* of $\boldsymbol{A}$, denoted as $\boldsymbol{A}'$, is the $k \times n$ matrix whose i th column is the i th row of $\boldsymbol{A}$. Clearly, a symmetric matrix is such that $\boldsymbol{A} = \boldsymbol{A}'$. Matrix transposition has the following properties:

1. $(\boldsymbol{A}')' = \boldsymbol{A}$;
2. $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$;
3. $(\boldsymbol{A}\boldsymbol{B})' = \boldsymbol{B}'\boldsymbol{A}'$.

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two n-dimensional column vectors, then $\boldsymbol{A}'$ and $\boldsymbol{B}'$ are row vectors so that $\boldsymbol{A}'\boldsymbol{B} = \boldsymbol{B}'\boldsymbol{A}$ is nothing but the inner product of $\boldsymbol{A}$ and $\boldsymbol{B}$. Note, however, that $\boldsymbol{A}'\boldsymbol{B} \neq \boldsymbol{A}\boldsymbol{B}'$, where $\boldsymbol{A}\boldsymbol{B}'$ is an $n \times n$ matrix, also known as the *outer product* of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Let $\boldsymbol{A}$ be $n \times k$ and $\boldsymbol{B}$ be $m \times r$. The *Kronecker product* of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the $nm \times kr$ matrix:

$$\boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1k}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2k}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}\boldsymbol{B} & a_{n2}\boldsymbol{B} & \cdots & a_{nk}\boldsymbol{B} \end{bmatrix}.$$

The Kronecker product has the following properties:

1. $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C}) \otimes (\boldsymbol{B}\boldsymbol{D})$;
2. $\boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes \boldsymbol{C}$;
3. $(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'$;
4. $(\boldsymbol{A} + \boldsymbol{B}) \otimes (\boldsymbol{C} + \boldsymbol{D}) = (\boldsymbol{A} \otimes \boldsymbol{C}) + (\boldsymbol{B} \otimes \boldsymbol{C}) + (\boldsymbol{A} \otimes \boldsymbol{D}) + (\boldsymbol{B} \otimes \boldsymbol{D})$.

It should be clear that the Kronecker product is *not* commutative: $\boldsymbol{A} \otimes \boldsymbol{B} \neq \boldsymbol{B} \otimes \boldsymbol{A}$. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $n \times k$ matrices, the *Hadamard product* (*direct product*) of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the matrix $\boldsymbol{A} * \boldsymbol{B} = \boldsymbol{C}$ with $c_{ij} = a_{ij}b_{ij}$.

3.3 Scalar Functions of Matrix Elements

Scalar functions of a matrix summarize various characteristics of matrix elements. An important scalar function is the *determinant* function. Formally, the determinant of a square matrix $\boldsymbol{A}$, denoted as $\det(\boldsymbol{A})$, is the sum of all signed elementary products from $\boldsymbol{A}$. To avoid excessive mathematical notations and definitions, we shall not discuss this definition; readers are referred to Anton (1981) and Basilevsky (1983) for more details. In what follows we first describe how determinants can be evaluated and then discuss their properties.

Given a square matrix $\boldsymbol{A}$, the *minor* of entry a_{ij}, denoted as M_{ij}, is the determinant of the submatrix that remains after the i th row and j th column are deleted from $\boldsymbol{A}$. The number $(-1)^{i+j}M_{ij}$ is called the *cofactor* of entry a_{ij}, denoted as C_{ij}. The determinant can be computed as follows; we omit the proof.

Theorem 3.1 *Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then for each $1 \leq i \leq n$ and $1 \leq j \leq n$,*

$$\det(\boldsymbol{A}) = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in},$$

where C_{ij} is the cofactor of a_{ij}.

This method is known as the *cofactor expansion* along the i th row. Similarly, the cofactor expansion along the j th column is

$$\det(\boldsymbol{A}) = a_{1j}C_{1j} + a_{2j}C_{2j} + \cdots + a_{nj}C_{nj}.$$

In particular, when $n = 2$, $\det(\boldsymbol{A}) = a_{11}a_{22} - a_{12}a_{21}$, a well known formula for 2×2 matrices. Clearly, if $\boldsymbol{A}$ contains a zero row or column, its determinant must be zero.

It is now not difficult to verify that the determinant of a diagonal or triangular matrix is the product of all elements on the main diagonal, i.e., $\det(\boldsymbol{A}) = \prod_i a_{ii}$. Let $\boldsymbol{A}$ be an $n \times n$ matrix and h a scalar. The determinant has the following properties:

1. $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$.

2. If $\boldsymbol{A}^*$ is obtained by multiplying a row (column) of $\boldsymbol{A}$ by h, then $\det(\boldsymbol{A}^*) = h\det(\boldsymbol{A})$.

3. If $\boldsymbol{A}^* = h\,\boldsymbol{A}$, then $\det(\boldsymbol{A}^*) = h^n\det(\boldsymbol{A})$.

4. If $\boldsymbol{A}^*$ is obtained by interchanging two rows (columns) of $\boldsymbol{A}$, then $\det(\boldsymbol{A}^*) = -\det(\boldsymbol{A})$.

5. If $\boldsymbol{A}^*$ is obtained by adding a scalar multiple of one row (column) to another row (column) of $\boldsymbol{A}$, then $\det(\boldsymbol{A}^*) = \det(\boldsymbol{A})$.

6. If a row (column) of $\boldsymbol{A}$ is linearly dependent on the other rows (columns), $\det(\boldsymbol{A}) = 0$.

7. $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.

8. $\det(\boldsymbol{A} \otimes \boldsymbol{B}) = \det(\boldsymbol{A})^n \det(\boldsymbol{B})^m$, where $\boldsymbol{A}$ is $n \times n$ and $\boldsymbol{B}$ is $m \times m$.

The *trace* of an $n \times n$ matrix $\boldsymbol{A}$ is the sum of diagonal elements, i.e., $\operatorname{trace}(\boldsymbol{A}) = \sum_{i=1}^n a_{ii}$. Clearly, the trace of the identity matrix is the number of its rows (columns). Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $n \times n$ matrices and h and k are two scalars. The trace function has the following properties:

1. $\operatorname{trace}(\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{A}')$;
2. $\operatorname{trace}(h\boldsymbol{A} + k\boldsymbol{B}) = h \operatorname{trace}(\boldsymbol{A}) + k \operatorname{trace}(\boldsymbol{B})$;
3. $\operatorname{trace}(\boldsymbol{AB}) = \operatorname{trace}(\boldsymbol{BA})$;
4. $\operatorname{trace}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{trace}(\boldsymbol{A}) \operatorname{trace}(\boldsymbol{B})$, where $\boldsymbol{A}$ is $n \times n$ and $\boldsymbol{B}$ is $m \times m$.

3.4 Matrix Rank

Given a matrix $\boldsymbol{A}$, the *row* (*column*) *space* is the space spanned by the row (column) vectors. The dimension of the row (column) space of $\boldsymbol{A}$ is called the *row rank* (*column rank*) of $\boldsymbol{A}$. Thus, the row (column) rank of a matrix is determined by the number of linearly independent row (column) vectors in that matrix. Suppose that $\boldsymbol{A}$ is an $n \times k$ ($k \leq n$) matrix with row rank $r \leq n$ and column rank $c \leq k$. Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$ be a basis for the row space. We can then write each row as

$$\boldsymbol{a}_i = b_{i1}\boldsymbol{u}_1 + b_{i2}\boldsymbol{u}_2 + \cdots + b_{ir}\boldsymbol{u}_r, \qquad i = 1, \ldots, n,$$

with the j th element

$$a_{ij} = b_{i1}u_{1j} + b_{i2}u_{2j} + \cdots + b_{ir}u_{rj}, \qquad i = 1, \ldots, n.$$

As $a_{1j}, \ldots, a_{nj}$ form the j th column of $\boldsymbol{A}$, each column of $\boldsymbol{A}$ is thus a linear combination of r vectors: $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_r$. It follows that the column rank of $\boldsymbol{A}$, c, is less than the row rank of $\boldsymbol{A}$, r. Similarly, the column rank of $\boldsymbol{A}'$ is less than the row rank of $\boldsymbol{A}'$. Note that the column (row) space of $\boldsymbol{A}'$ is nothing but the row (column) space of $\boldsymbol{A}$. Thus, the row rank of $\boldsymbol{A}$, r, is less than the column rank of $\boldsymbol{A}$, c. Combining these results we have the following conclusion.

Theorem 3.2 *The row rank and column rank of a matrix are equal.*

In the light of Theorem 3.2, the *rank* of a matrix $\boldsymbol{A}$ is defined to be the dimension of the row (or column) space of $\boldsymbol{A}$, denoted as $\operatorname{rank}(\boldsymbol{A})$. A square $n \times n$ matrix $\boldsymbol{A}$ is said to be of *full rank* if $\operatorname{rank}(\boldsymbol{A}) = n$.

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $n \times k$ matrices and $\boldsymbol{C}$ be a $k \times m$ matrix. The rank has the following properties:

1. $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}')$;
2. $\operatorname{rank}(\boldsymbol{A} + \boldsymbol{B}) \leq \operatorname{rank}(\boldsymbol{A}) + \operatorname{rank}(\boldsymbol{B})$;
3. $\operatorname{rank}(\boldsymbol{A} - \boldsymbol{B}) \geq |\operatorname{rank}(\boldsymbol{A}) - \operatorname{rank}(\boldsymbol{B})|$;
4. $\operatorname{rank}(\boldsymbol{A}\boldsymbol{C}) \leq \min(\operatorname{rank}(\boldsymbol{A}), \operatorname{rank}(C))$;
5. $\operatorname{rank}(\boldsymbol{A}\boldsymbol{C}) \geq \operatorname{rank}(\boldsymbol{A}) + \operatorname{rank}(\boldsymbol{C}) - k$;
6. $\operatorname{rank}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{rank}(\boldsymbol{A}) \operatorname{rank}(\boldsymbol{B})$, where $\boldsymbol{A}$ is $m \times n$ and $\boldsymbol{B}$ is $p \times q$.

3.5 Matrix Inversion

For any square matrix $\boldsymbol{A}$, if there exists a matrix $\boldsymbol{B}$ such that $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}$, then $\boldsymbol{A}$ is said to be *invertible*, and $\boldsymbol{B}$ is the *inverse* of $\boldsymbol{A}$, denoted as $\boldsymbol{A}^{-1}$. Note that $\boldsymbol{A}^{-1}$ need not exist; if it does, it is unique. (Verify!) An invertible matrix is also known as a *nonsingular matrix*; otherwise, it is *singular*. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two nonsingular matrices. Matrix inversion has the following properties:

1. $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$;
2. $(\boldsymbol{A}^{-1})' = (\boldsymbol{A}')^{-1}$;
3. $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$;
4. $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$;
5. $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$;
6. $(\boldsymbol{A}^{r})^{-1} = (\boldsymbol{A}^{-1})^{r}$.

An $n \times n$ matrix is an *elementary matrix* if it can be obtained from the $n \times n$ identity matrix $\boldsymbol{I}_n$ by a single *elementary operation*. By elementary row (column) operations we mean:

- Multiply a row (column) by a non-zero constant.
- Interchange two rows (columns).
- Add a scalar multiple of one row (column) to another row (column).

The matrices below are examples of elementary matrices:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

It can be verified that pre-multiplying (post-multiplying) a matrix $\boldsymbol{A}$ by an elementary matrix is equivalent to performing an elementary row (column) operation on $\boldsymbol{A}$. For an elementary matrix $\boldsymbol{E}$, there is an *inverse operation*, which is also an elementary operation, on $\boldsymbol{E}$ to produce $\boldsymbol{I}$. Let $\boldsymbol{E}^*$ denote the elementary matrix associated with the inverse operation. Then, $\boldsymbol{E}^*\boldsymbol{E} = \boldsymbol{I}$. Likewise, $\boldsymbol{E}\boldsymbol{E}^* = \boldsymbol{I}$. This shows that $\boldsymbol{E}$ is invertible and $\boldsymbol{E}^* = \boldsymbol{E}^{-1}$. Matrices that can be obtained from one another by finitely many elementary row (column) operations are said to be *row (column) equivalent.* The result below is useful in practice; we omit the proof.

Theorem 3.3 *Let $\boldsymbol{A}$ be a square matrix. The following statements are equivalent.*

(a) *$\boldsymbol{A}$ is invertible (nonsingular).*

(b) *$\boldsymbol{A}$ is row (column) equivalent to $\boldsymbol{I}$.*

(c) $\det(\boldsymbol{A}) \neq 0$.

(d) *$\boldsymbol{A}$ is of full rank.*

Let $\boldsymbol{A}$ be a nonsingular $n \times n$ matrix, $\boldsymbol{B}$ an $n \times k$ matrix, and $\boldsymbol{C}$ an $n \times n$ matrix. It is easy to verify that $\operatorname{rank}(\boldsymbol{B}) = \operatorname{rank}(\boldsymbol{AB})$ and $\operatorname{trace}(\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{C})$.

When $\boldsymbol{A}$ is row equivalent to $\boldsymbol{I}$, then there exist elementary matrices $\boldsymbol{E}_1, \ldots, \boldsymbol{E}_r$ such that $\boldsymbol{E}_r \cdots \boldsymbol{E}_2 \boldsymbol{E}_1 \boldsymbol{A} = \boldsymbol{I}$. As elementary matrices are invertible, pre-multiplying their inverses on both sides yields

$$\boldsymbol{A} = \boldsymbol{E}_1^{-1}\boldsymbol{E}_2^{-1} \cdots \boldsymbol{E}_r^{-1}, \qquad \boldsymbol{A}^{-1} = \boldsymbol{E}_r \cdots \boldsymbol{E}_2 \boldsymbol{E}_1 \boldsymbol{I}.$$

Thus, the inverse of a matrix $\boldsymbol{A}$ can be obtained by performing finitely many elementary row (column) operations on the augmented matrix $[\boldsymbol{A} : \boldsymbol{I}]$:

$$\boldsymbol{E}_r \cdots \boldsymbol{E}_2 \boldsymbol{E}_1 [\boldsymbol{A} : \boldsymbol{I}] = [\boldsymbol{I} : \boldsymbol{A}^{-1}].$$

Given an $n \times n$ matrix $\boldsymbol{A}$, let C_{ij} be the cofactor of a_{ij}. The transpose of the matrix of cofactors, (C_{ij}), is called the *adjoint* of $\boldsymbol{A}$, denoted as $\operatorname{adj}(\boldsymbol{A})$. The inverse of a matrix can also be computed using its adjoint matrix, as shown in the following result; for a proof see Anton (1981, p. 81).

Theorem 3.4 *Let* $\boldsymbol{A}$ *be an invertible matrix, then*

$$\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})} \operatorname{adj}(\boldsymbol{A}).$$

For a rectangular $n \times k$ matrix $\boldsymbol{A}$, the *left inverse* of $\boldsymbol{A}$ is a $k \times n$ matrix $\boldsymbol{A}_L^{-1}$ such that $\boldsymbol{A}_L^{-1}\boldsymbol{A} = \boldsymbol{I}_k$, and the *right inverse* of $\boldsymbol{A}$ is a $k \times n$ matrix $\boldsymbol{A}_R^{-1}$ such that $\boldsymbol{A}\boldsymbol{A}_R^{-1} = \boldsymbol{I}_n$. The left inverse and right inverse are *not* unique, however. Let $\boldsymbol{A}$ be an $n \times k$ matrix. We now present an important result without a proof.

Theorem 3.5 *Given an* $n \times k$ *matrix,*

(a) $\boldsymbol{A}$ *has a left inverse if, and only if,* $\operatorname{rank}(\boldsymbol{A}) = k \leq n$*, and*

(b) $\boldsymbol{A}$ *has a right inverse if, and only if,* $\operatorname{rank}(\boldsymbol{A}) = n \leq k$*.*

If an $n \times k$ matrix $\boldsymbol{A}$ has both left and right inverses, then it must be the case that $k = n$ and

$$\boldsymbol{A}_L^{-1} = \boldsymbol{A}_L^{-1}\boldsymbol{A}\boldsymbol{A}_R^{-1} = \boldsymbol{A}_R^{-1}.$$

Hence, $\boldsymbol{A}_L^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}_L^{-1} = \boldsymbol{I}$ so that $\boldsymbol{A}_R^{-1} = \boldsymbol{A}_L^{-1} = \boldsymbol{A}^{-1}$.

Theorem 3.6 *If* $\boldsymbol{A}$ *has both left and right inverses, then* $\boldsymbol{A}_L^{-1} = \boldsymbol{A}_R^{-1} = A^{-1}$*.*

3.6 Statistical Applications

Given an $n \times 1$ matrix of random variables $\boldsymbol{x}$, $\mathbb{E}(\boldsymbol{x})$ is the $n \times 1$ matrix containing the expectations of x_i. Let $\boldsymbol{\ell}$ denote the vector of ones. The variance-covariance matrix of $\boldsymbol{x}$ is the expectation of the outer product of $\boldsymbol{x} - \mathbb{E}(\boldsymbol{x})$:

$$\begin{aligned}
\operatorname{var}(\boldsymbol{x}) &= \mathbb{E}[(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))'] \\
&= \begin{bmatrix} \operatorname{var}(x_1) & \operatorname{cov}(x_1, x_2) & \cdots & \operatorname{cov}(x_1, x_n) \\ \operatorname{cov}(x_2, x_1) & \operatorname{var}(x_2) & \cdots & \operatorname{cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(x_n, x_1) & \operatorname{cov}(x_n, x_2) & \cdots & \operatorname{var}(x_n) \end{bmatrix}.
\end{aligned}$$

As $\operatorname{cov}(x_i, x_j) = \operatorname{cov}(x_j, x_i)$, $\operatorname{var}(\boldsymbol{x})$ must be symmetric. If $\operatorname{cov}(x_i, x_j) = 0$ for all $i \neq j$ so that x_i and x_j are uncorrelated, then $\operatorname{var}(\boldsymbol{x})$ is diagonal. When a component of $\boldsymbol{x}$, say x_i, is non-stochastic, we have $\operatorname{var}(x_i) = 0$ and $\operatorname{cov}(x_i, x_j) = 0$ for all $j \neq i$. In this case,

$\det(\text{var}(\boldsymbol{x})) = 0$ by Theorem 3.1, so that $\text{var}(\boldsymbol{x})$ is singular. It is straightforward to verify that for a square matrix $\boldsymbol{A}$, $\mathbb{E}[\text{trace}(\boldsymbol{A})] = \text{trace}(\mathbb{E}[\boldsymbol{A}])$, but $\mathbb{E}[\det(\boldsymbol{A})] \neq \det(\mathbb{E}[\boldsymbol{A}])$.

Let h be a non-stochastic scalar and $\boldsymbol{x}$ an $n \times 1$ matrix (vector) of random variables. We have $\mathbb{E}(h\boldsymbol{x}) = h\,\mathbb{E}(\boldsymbol{x})$ and for $\boldsymbol{\Sigma} = \text{var}(\boldsymbol{x})$,

$$\text{var}(h\boldsymbol{x}) = \mathbb{E}[h^2(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))'] = h^2\boldsymbol{\Sigma}.$$

For a non-stochastic $n \times 1$ matrix $\boldsymbol{a}$, $\boldsymbol{a}'\boldsymbol{x}$ is a linear combination of the random variables in $\boldsymbol{x}$, and hence a random variable. We then have $\mathbb{E}(\boldsymbol{a}'\boldsymbol{x}) = \boldsymbol{a}'\,\mathbb{E}(\boldsymbol{x})$ and

$$\text{var}(\boldsymbol{a}'\boldsymbol{x}) = \mathbb{E}[\boldsymbol{a}'(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))'\boldsymbol{a}] = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a},$$

which are scalars. Similarly, for a non-stochastic $m \times n$ matrix $\boldsymbol{A}$, $\mathbb{E}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{x})$ and $\text{var}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$. Note that when $\boldsymbol{A}$ is not of full row rank, $\text{rank}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}') = \text{rank}(\boldsymbol{A}) < m$. Thus, $\boldsymbol{A}\boldsymbol{x}$ is *degenerate* in the sense that $\text{var}(\boldsymbol{A}\boldsymbol{x})$ is a singular matrix.

Exercises

3.1 Consider two matrices:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 4 & 5 \\ 1 & 1 & 2 \\ 0 & 6 & 8 \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 2 & 2 & 9 \\ 6 & 1 & 6 \\ 5 & 1 & 0 \end{bmatrix}.$$

Show that $\boldsymbol{AB} \neq \boldsymbol{BA}$.

3.2 Find two non-zero matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that $\boldsymbol{AB} = \boldsymbol{o}$.

3.3 Find three matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ such that $\boldsymbol{AB} = \boldsymbol{AC}$ but $\boldsymbol{B} \neq \boldsymbol{C}$.

3.4 Let $\boldsymbol{A}$ be a 3×3 matrix. Apply the cofactor expansion along the first row of $\boldsymbol{A}$ to obtain a formula for $\det(\boldsymbol{A})$.

3.5 Let $\boldsymbol{X}$ be an $n \times k$ matrix with rank $k < n$. Find $\text{trace}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$.

3.6 Let $\boldsymbol{X}$ be an $n \times k$ matrix with $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X}'\boldsymbol{X}) = k < n$. Find

$$\text{rank}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}').$$

3.7 Let $\boldsymbol{A}$ be a symmetric matrix. Show that when $\boldsymbol{A}$ is invertible, $\boldsymbol{A}^{-1}$ and $\text{adj}(\boldsymbol{A})$ are also symmetric.

3.8 Apply Theorem 3.4 to find the inverse of

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

4 Linear Transformation

There are two ways to relocate a vector $\boldsymbol{v}$: transforming $\boldsymbol{v}$ to a new position and transforming the associated coordinate axes or basis.

4.1 Change of Basis

Let $B = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$ be a basis for a vector space V. Then $\boldsymbol{v} \in V$ can be written as

$$\boldsymbol{v} = c_1 \boldsymbol{u}_1 + c_2 \boldsymbol{u}_2 + \cdots + c_k \boldsymbol{u}_k.$$

The coefficients $c_1, \ldots, c_k$ form the *coordinate vector* of $\boldsymbol{v}$ relative to B. We write $[\boldsymbol{v}]_B$ as the coordinate column vector relative to B. Clearly, if B is the collection of Cartesian unit vectors, then $[\boldsymbol{v}]_B = \boldsymbol{v}$.

Consider now a two-dimensional space with an old basis $B = \{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ and a new basis $B' = \{\boldsymbol{w}_1, \boldsymbol{w}_2\}$. Then,

$$\boldsymbol{u}_1 = a_1 \boldsymbol{w}_1 + a_2 \boldsymbol{w}_2, \qquad \boldsymbol{u}_2 = b_1 \boldsymbol{w}_1 + b_2 \boldsymbol{w}_2,$$

so that

$$[\boldsymbol{u}_1]_{B'} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \qquad [\boldsymbol{u}_2]_{B'} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

For $\boldsymbol{v} = c_1 \boldsymbol{u}_1 + c_2 \boldsymbol{u}_2$, we can write this vector in terms of the new basis vectors as

$$\begin{aligned} \boldsymbol{v} &= c_1(a_1 \boldsymbol{w}_1 + a_2 \boldsymbol{w}_2) + c_2(b_1 \boldsymbol{w}_1 + b_2 \boldsymbol{w}_2) \\ &= (c_1 a_1 + c_2 b_1)\boldsymbol{w}_1 + (c_1 a_2 + c_2 b_2)\boldsymbol{w}_2. \end{aligned}$$

It follows that

$$[\boldsymbol{v}]_{B'} = \begin{bmatrix} c_1 a_1 + c_2 b_1 \\ c_1 a_2 + c_2 b_2 \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \Big[[\boldsymbol{u}_1]_{B'} : [\boldsymbol{u}_2]_{B'} \Big] [\boldsymbol{v}]_B,$$

where $\boldsymbol{P}$ is called the *transition matrix* from B to B'. That is, the new coordinate vector can be obtained by pre-multiplying the old coordinate vector by the transition matrix P, where the column vectors of P are the coordinate vectors of the old basis vectors relative to the new basis.

More generally, let $B = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$ and $B' = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}$. Then $[\boldsymbol{v}]_{B'} = \boldsymbol{P}\,[\boldsymbol{v}]_B$ with the transition matrix

$$\boldsymbol{P} = \Big[[\boldsymbol{u}_1]_{B'} : [\boldsymbol{u}_2]_{B'} : \ \cdots \ : [\boldsymbol{u}_k]_{B'} \Big].$$

The following examples illustrate the properties of P.

Examples:

1. Consider two bases: $B = \{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ and $B' = \{\boldsymbol{w}_1, \boldsymbol{w}_2\}$, where

$$\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \boldsymbol{u}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \boldsymbol{w}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

As $\boldsymbol{u}_1 = -\boldsymbol{w}_1 + \boldsymbol{w}_2$ and $\boldsymbol{u}_2 = 2\boldsymbol{w}_1 - \boldsymbol{w}_2$, we have

$$\boldsymbol{P} = \begin{bmatrix} -1 & 2 \\ 1 & -1 \end{bmatrix}.$$

If

$$\boldsymbol{v} = \begin{bmatrix} 7 \\ 2 \end{bmatrix},$$

then $[\boldsymbol{v}]_B = \boldsymbol{v}$ and

$$[\boldsymbol{v}]_{B'} = \boldsymbol{P}\boldsymbol{v} = \begin{bmatrix} -3 \\ 5 \end{bmatrix}.$$

Similarly, we can show the transition matrix from B' to B is

$$\boldsymbol{Q} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}.$$

It is interesting to note that $\boldsymbol{PQ} = \boldsymbol{QP} = \boldsymbol{I}$ so that $\boldsymbol{Q} = \boldsymbol{P}^{-1}$.

2. Rotation of the standard xy-coordinate system counterclockwise about the origin through an angle θ to a new $x'y'$-system such that the angle between basis vectors and vector lengths are preserved. Let $B = \{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ and $B' = \{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ be the bases for the old and new systems, respectively, where $\boldsymbol{u}$ and $\boldsymbol{w}$ are associated unit vectors. Note that

$$[\boldsymbol{u}_1]_{B'} = \begin{bmatrix} \cos\theta \\ -\sin\theta \end{bmatrix}, \qquad [\boldsymbol{u}_2]_{B'} = \begin{bmatrix} \cos(\pi/2 - \theta) \\ \sin(\pi/2 - \theta) \end{bmatrix} = \begin{bmatrix} \sin\theta \\ \cos\theta \end{bmatrix}.$$

Thus,

$$\boldsymbol{P} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Similarly, the transition matrix from B' to B is

$$\boldsymbol{P}^{-1} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

It is also interesting to note that $\boldsymbol{P}^{-1} = \boldsymbol{P}'$; this property does not hold in the first example, however.

More generally, we have the following result; the proof is omitted.

Theorem 4.1 *If P is the transition matrix from a basis B to a new basis B', then P is invertible and P^{-1} is the transition matrix from B' to B. If both B and B' are orthonormal, then $P^{-1} = P'$.*

4.2 Systems of Linear Equations

Let $\boldsymbol{y}$ be an $n \times 1$ matrix, $\boldsymbol{x}$ an $m \times 1$ matrix, and $\boldsymbol{A}$ an $n \times m$ matrix. Then, $\boldsymbol{Ax} = \boldsymbol{y}$ ($\boldsymbol{Ax} = \boldsymbol{o}$) represents a *non-homogeneous* (*homogeneous*) system of n linear equations with m unknowns. A system of equations is said to be *inconsistent* if it has no solution, and a system is *consistent* if it has at least one solution. A non-homogeneous system has either no solution, exactly one solution or infinitely many solutions. On the other hand, a homogeneous system has either the trivial solution (i.e., $\boldsymbol{x} = 0$) or infinitely many solutions (including the trivial solution), and hence must be consistent.

Consider the following examples:

1. No solution: The system

$$\begin{aligned} x_1 + x_2 &= 4, \\ 2x_1 + 2x_2 &= 7, \end{aligned}$$

involves two parallel lines, and hence has no solution.

2. Exactly one solution: The system

$$\begin{aligned} x_1 + x_2 &= 4, \\ 2x_1 + 3x_2 &= 8, \end{aligned}$$

has one solution: $x_1 = 4$, $x_2 = 0$. That is, these two lines intersect at $(4, 0)$.

3. Infinitely many solutions: The system

$$\begin{aligned} x_1 + x_2 &= 4, \\ 2x_1 + 2x_2 &= 8, \end{aligned}$$

yields only one line, and hence has infinitely many solutions.

4.3 Linear Transformation

Let V and W be two vector spaces and $T : V \to W$ is a function mapping V into W. T is said to be a *linear transformation* if $T(\boldsymbol{u} + \boldsymbol{v}) = T(\boldsymbol{u}) + T(\boldsymbol{v})$ for all $\boldsymbol{u}, \boldsymbol{v}$ in V and $T(h\boldsymbol{u}) = hT(\boldsymbol{u})$ for all $\boldsymbol{u}$ in V and all scalars h. Note that $T(\boldsymbol{o}) = \boldsymbol{o}$ and $T(-\boldsymbol{u}) = -T(\boldsymbol{u})$. Let $\boldsymbol{A}$ be a fixed $n \times m$ matrix. Then $T : \Re^m \to \Re^n$ such that $T(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$ is a linear transformation.

Let $T\colon \Re^2 \to \Re^2$ denote the multiplication by a 2×2 matrix $\boldsymbol{A}$.

1. Identity transformation—mapping each point into itself: $\boldsymbol{A} = \boldsymbol{I}$.

2. Reflection about the y-axis—mapping (x, y) to $(-x, y)$:

$$\boldsymbol{A} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

3. Reflection about the x-axis—mapping (x, y) to $(x, -y)$:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

4. Reflection about the line $x = y$— mapping (x, y) to (y, x):

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

5. Rotation counterclockwise through an angle θ:

$$\boldsymbol{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Let ϕ denote the angle between the vector $\boldsymbol{v} = (x, y)$ and the positive x-axis and ν denote the norm of $\boldsymbol{v}$. Then $x = \nu \cos\phi$ and $y = \nu \sin\phi$, and

$$\begin{aligned} \boldsymbol{A}\boldsymbol{v} &= \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \nu\cos\phi\cos\theta - \nu\sin\phi\sin\theta \\ \nu\cos\phi\sin\theta + \nu\sin\phi\cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \nu\cos(\phi+\theta) \\ \nu\sin(\phi+\theta) \end{bmatrix}. \end{aligned}$$

Rotation through an angle $-\theta$ can then be obtained by multiplication of

$$\boldsymbol{A} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Note that the orthogonal projection and the mapping of a vector into its coordinate vector with respect to a basis are also linear transformations.

If $T : V \to W$ is a linear transformation, the set $\{\boldsymbol{v} \in V : T(\boldsymbol{v}) = \boldsymbol{o}\}$ is called the *kernel* or *null space* of T, denoted as $\ker(T)$. The set $\{\boldsymbol{w} \in W : T(\boldsymbol{v}) = \boldsymbol{w}$ for some $\boldsymbol{v} \in V\}$ is the *range* of T, denoted as $\mathrm{range}(T)$. Clearly, the kernel and range of T are both closed under vector addition and scalar multiplication, and hence are subspaces of V and W, respectively. The dimension of the range of T is called the *rank* of T, and the dimension of the kernel is called the *nullity* of T.

Let $T : \Re^m \to \Re^n$ denote multiplication by an $n \times m$ matrix $\boldsymbol{A}$. Note that $\boldsymbol{y}$ is a linear combination of the column vectors of $\boldsymbol{A}$ if, and only if, it can be expressed as the consistent linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$:

$$x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_m\boldsymbol{a}_m = \boldsymbol{y},$$

where $\boldsymbol{a}_j$ is the j^{th} column of $\boldsymbol{A}$. Thus, $\mathrm{range}(T)$ is also the column space of $\boldsymbol{A}$, and the rank of T is the column rank of $\boldsymbol{A}$. The linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is therefore a linear transformation of $\boldsymbol{x} \in V$ to some $\boldsymbol{y}$ in the column space of $\boldsymbol{A}$. It is also clear that $\ker(T)$ is the solution space of the homogeneous system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$.

Theorem 4.2 *If $T : V \to W$ is a linear transformation, where V is an m-dimensional space, then*

$$\dim(\mathrm{range}(T)) + \dim(\ker(T)) = m,$$

i.e., $\mathrm{rank}(T) + \text{nullity of } T = m$.

Proof: Suppose first that the kernel of T has the dimension $1 \le \dim(\ker(T)) = r < m$ and a basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$. Then by Theorem 2.3, this set can be enlarged such that $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r, \boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_m\}$ is a basis for V. Let $\boldsymbol{w}$ be an arbitrary vector in range(T), then $\boldsymbol{w} = T(\boldsymbol{u})$ for some $\boldsymbol{u} \in V$, where $\boldsymbol{u}$ can be written as

$$\boldsymbol{u} = c_1\boldsymbol{v}_1 + \cdots + c_r\boldsymbol{v}_r + c_{r+1}\boldsymbol{v}_{r+1} + \cdots + c_m\boldsymbol{v}_m.$$

As $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$ is in the kernel, $T(\boldsymbol{v}_1) = \cdots = T(\boldsymbol{v}_r) = \boldsymbol{o}$ so that

$$\boldsymbol{w} = T(\boldsymbol{u}) = c_{r+1}T(\boldsymbol{v}_{r+1}) + \cdots + c_mT(\boldsymbol{v}_m).$$

This shows that $S = \{T(\boldsymbol{v}_{r+1}), \ldots, T(\boldsymbol{v}_m)\}$ spans range(T). If we can show that S is an independent set, then S is a basis for range(T), and consequently,

$$\dim(\mathrm{range}(T)) + \dim(\ker(T)) = (m-r) + r = m.$$

Observe that for

$$\boldsymbol{o} = h_{r+1}T(\boldsymbol{v}_{r+1}) + \cdots + h_mT(\boldsymbol{v}_m) = T(h_{r+1}\boldsymbol{v}_{r+1} + \cdots + h_m\boldsymbol{v}_m),$$

$h_{r+1}\boldsymbol{v}_{r+1} + \cdots + h_m\boldsymbol{v}_m$ is in the kernel of T. Then for some $h_1, \ldots, h_r$,

$$h_{r+1}\boldsymbol{v}_{r+1} + \cdots + h_m\boldsymbol{v}_m = h_1\boldsymbol{v}_1 + \cdots + h_r\boldsymbol{v}_r.$$

As $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r, \boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_m\}$ is a basis for V, all h's of this equation must be zero. This proves that S is an independent set. We now prove the assertion for $\dim(\ker(T)) = m$. In this case, $\ker(T)$ must be V, and for every $\boldsymbol{u} \in V$, $T(\boldsymbol{u}) = \boldsymbol{o}$. That is, range$(T) = \{\boldsymbol{o}\}$. The proof for the case $\dim(\ker(T)) = 0$ is left as an exercise. $\square$

The next result follows straightforwardly from Theorem 4.2.

Corollary 4.3 *Let $\boldsymbol{A}$ be an $n \times m$ matrix. The dimension of the solution space of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$ is $m - \mathrm{rank}(\boldsymbol{A})$.*

Hence, given an $n \times m$ matrix, the homogeneous system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$ has the trivial solution if $\mathrm{rank}(\boldsymbol{A}) = m$. When $\boldsymbol{A}$ is square, $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$ has the trivial solution if, and only if, $\boldsymbol{A}$ is nonsingular; when $\boldsymbol{A}$ has rank $m < n$, $\boldsymbol{A}$ has a left inverse by Theorem 3.5 so that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$ also has the trivial solution. This system has infinitely many solutions if $\boldsymbol{A}$ is not of full column rank. This is the case when the number of unknowns is greater than the number of equations ($\mathrm{rank}(\boldsymbol{A}) \le n < m$).

Theorem 4.4 *Let $\boldsymbol{A}^* = [\boldsymbol{A} : \boldsymbol{y}]$. The non-homogeneous system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is consistent if, and only if, $\mathrm{rank}(\boldsymbol{A}) = \mathrm{rank}(\boldsymbol{A}^*)$.*

Proof: Let $\boldsymbol{x}^*$ be the $(m+1)$-dimensional vector containing $\boldsymbol{x}$ and -1. If $\boldsymbol{Ax} = \boldsymbol{y}$ has a solution, $\boldsymbol{A}^*\boldsymbol{x}^* = \boldsymbol{o}$ has a non-trivial solution so that $\boldsymbol{A}^*$ is not of full column rank. Since $\operatorname{rank}(\boldsymbol{A}) \leq \operatorname{rank}(\boldsymbol{A}^*)$, we must have $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}^*)$. Conversely, if $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}^*)$, then by the definition of $\boldsymbol{A}^*$, $\boldsymbol{y}$ must be in the column space of $\boldsymbol{A}$. Thus, $\boldsymbol{Ax} = \boldsymbol{y}$ has a solution. □

For an $n \times m$ matrix $\boldsymbol{A}$, the non-homogeneous system $\boldsymbol{Ax} = \boldsymbol{y}$ has a unique solution if, and only if, $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}^*) = m$. If $\boldsymbol{A}$ is square, $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$ is of course unique. If $\boldsymbol{A}$ is rectangular with rank $m < n$, $\boldsymbol{x} = \boldsymbol{A}_L^{-1}\boldsymbol{y}$. Given that A_L is not unique, suppose that there are two solutions $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. We have

$$\boldsymbol{A}(\boldsymbol{x}_1 - \boldsymbol{x}_2) = \boldsymbol{y} - \boldsymbol{y} = \boldsymbol{o}.$$

Hence, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ must coincide, and the solution is also unique. If $\boldsymbol{A}$ is rectangular with rank $n < m$, $\boldsymbol{x}_0 = \boldsymbol{A}_R^{-1}\boldsymbol{y}$ is clearly a solution. In contrast with the previous case, this solution is not unique because $\boldsymbol{A}$ is not of full column rank so that $\boldsymbol{A}(\boldsymbol{x}_1 - \boldsymbol{x}_2) = \boldsymbol{o}$ has infinitely many solutions. If $\operatorname{rank}(\boldsymbol{A}) < \operatorname{rank}(\boldsymbol{A}^*)$, the system is inconsistent.

If linear transformations are performed in succession using matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$, they are equivalent to the transformation based on a single matrix $\boldsymbol{A} = \boldsymbol{A}_k\boldsymbol{A}_{k-1}\cdots\boldsymbol{A}_1$, i.e., $\boldsymbol{A}_k\boldsymbol{A}_{k-1}\cdots\boldsymbol{A}_1\boldsymbol{x} = \boldsymbol{Ax}$. As $\boldsymbol{A}_1\boldsymbol{A}_2 \neq \boldsymbol{A}_2\boldsymbol{A}_1$ in general, changing the order of $\boldsymbol{A}_i$ matrices results in different transformations. We also note that $\boldsymbol{Ax} = \boldsymbol{o}$ if, and only if, $\boldsymbol{x}$ is orthogonal to every row vector of $\boldsymbol{A}$, or equivalently, every column vector of $\boldsymbol{A}'$. This shows that the null space of $\boldsymbol{A}$ and the range space of $\boldsymbol{A}'$ are orthogonal.

Exercises

4.1 Let $T\colon \Re^4 \to \Re^3$ denote multiplication by

$$\begin{bmatrix} 4 & 1 & -2 & -3 \\ 2 & 1 & 1 & -4 \\ 6 & 0 & -9 & 9 \end{bmatrix}.$$

Which of the following vectors are in $\operatorname{range}(T)$ or $\ker(T)$?

$$\begin{bmatrix} 0 \\ 0 \\ 6 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ -8 \\ 2 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ -4 \\ 1 \\ 0 \end{bmatrix}.$$

4.2 Find the rank and nullity of $T\colon \Re^n \to \Re^n$ defined by: (i) $T(\boldsymbol{u}) = \boldsymbol{u}$, (ii) $T(\boldsymbol{u}) = \boldsymbol{o}$, (iii) $T(\boldsymbol{u}) = 3\boldsymbol{u}$.

4.3 Each of the following matrices transforms a point (x, y) to a new point (x', y'):

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 1/2 & 1 \end{bmatrix}.$$

Draw figures to show these (x, y) and (x', y').

4.4 Consider the vector $\boldsymbol{u} = (2, 1)$ and matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \qquad \boldsymbol{A}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Find $\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{u}$ and $\boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{u}$ and draw a figure to illustrate the transformed points.

5 Special Matrices

In what follows, we treat an n-dimensional vector as an $n \times 1$ matrix and use these terms interchangeably. We also let $\mathrm{span}(\boldsymbol{A})$ denote the column space of $\boldsymbol{A}$. Thus, $\mathrm{span}(\boldsymbol{A}')$ is the row space of $\boldsymbol{A}$.

5.1 Symmetric Matrix

We have learned that a matrix $\boldsymbol{A}$ is symmetric if $\boldsymbol{A} = \boldsymbol{A}'$. Let $\boldsymbol{A}$ be an $n \times k$ matrix. Then, $\boldsymbol{A}'\boldsymbol{A}$ is a $k \times k$ symmetric matrix with the (i, j) thelement $\boldsymbol{a}_i'\boldsymbol{a}_j$, where $\boldsymbol{a}_i$ is the i thcolumn of $\boldsymbol{A}$. If $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{o}$, then all the main diagonal elements are $\boldsymbol{a}_i'\boldsymbol{a}_i = \|\boldsymbol{a}_i\|^2 = 0$. It follows that all columns of $\boldsymbol{A}$ are zero vectors and hence $\boldsymbol{A} = \boldsymbol{o}$.

As $\boldsymbol{A}'\boldsymbol{A}$ is symmetric, its row space and column space are the same. If $\boldsymbol{x} \in \mathrm{span}(\boldsymbol{A}'\boldsymbol{A})^{\perp}$, i.e., $(\boldsymbol{A}'\boldsymbol{A})\boldsymbol{x} = \boldsymbol{o}$, then $\boldsymbol{x}'(\boldsymbol{A}'\boldsymbol{A})\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})'(\boldsymbol{A}\boldsymbol{x}) = 0$ so that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$. That is, $\boldsymbol{x}$ is orthogonal to every row vector of $\boldsymbol{A}$, and $\boldsymbol{x} \in \mathrm{span}(\boldsymbol{A}')^{\perp}$. This shows $\mathrm{span}(\boldsymbol{A}'\boldsymbol{A})^{\perp} \subseteq \mathrm{span}(\boldsymbol{A}')^{\perp}$. Conversely, if $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$, then $(\boldsymbol{A}'\boldsymbol{A})\boldsymbol{x} = \boldsymbol{o}$. This shows $\mathrm{span}(\boldsymbol{A}')^{\perp} \subseteq \mathrm{span}(\boldsymbol{A}'\boldsymbol{A})^{\perp}$. We have established:

Theorem 5.1 *Let $\boldsymbol{A}$ be an $n \times k$ matrix. Then the row space of $\boldsymbol{A}$ is the same as the row space of $\boldsymbol{A}'\boldsymbol{A}$.*

Similarly, the column space of $\boldsymbol{A}$ is the same as the column space of $\boldsymbol{A}\boldsymbol{A}'$. It follows from Theorem 3.2 and Theorem 5.1 that:

Theorem 5.2 *Let $\boldsymbol{A}$ be an $n \times k$ matrix. Then* $\mathrm{rank}(\boldsymbol{A}) = \mathrm{rank}(\boldsymbol{A}'\boldsymbol{A}) = \mathrm{rank}(\boldsymbol{A}\boldsymbol{A}')$.

In particular, if $\boldsymbol{A}$ is of rank $k < n$, then $\boldsymbol{A}'\boldsymbol{A}$ is of full rank k so that $\boldsymbol{A}'\boldsymbol{A}$ is nonsingular, but $\boldsymbol{A}\boldsymbol{A}'$ is not of full rank, and hence a singular matrix.

5.2 Skew-Symmetric Matrix

A square matrix $\boldsymbol{A}$ is said to be *skew symmetric* if $\boldsymbol{A} = -\boldsymbol{A}'$. Note that for any square matrix $\boldsymbol{A}$, $\boldsymbol{A} - \boldsymbol{A}'$ is skew-symmetric and $\boldsymbol{A} + \boldsymbol{A}'$ is symmetric. Thus, any matrix $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \frac{1}{2}(\boldsymbol{A} - \boldsymbol{A}') + \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}').$$

As the main diagonal elements are not altered by transposition, a skew-symmetric matrix $\boldsymbol{A}$ must have zeros on the main diagonal so that $\mathrm{trace}(\boldsymbol{A}) = 0$. It is also easy to verify that the sum of two skew-symmetric matrices is skew-symmetric and that the square of a skew-symmetric (or symmetric) matrix is symmetric because

$$\boldsymbol{A}^2 = (\boldsymbol{A}'\boldsymbol{A}')' = ((-\boldsymbol{A})(-\boldsymbol{A}))' = (\boldsymbol{A}^2)'.$$

An interesting property of an $n \times n$ skew-symmetric matrix $\boldsymbol{A}$ is that $\boldsymbol{A}$ is singular if n is odd. To see this, note that

$$\det(\boldsymbol{A}) = \det(-\boldsymbol{A}') = (-1)^n \det(\boldsymbol{A}).$$

When n is odd, $\det(\boldsymbol{A}) = -\det(\boldsymbol{A})$ so that $\det(\boldsymbol{A}) = 0$. By Theorem 3.3, $\boldsymbol{A}$ is singular.

5.3 Quadratic Form and Definite Matrix

Recall that a second order polynomial in the variables $x_1, \ldots, x_n$ is

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j,$$

which can be expressed as a *quadratic form*: $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}$ is $n \times 1$ and $\boldsymbol{A}$ is $n \times n$. We know that an arbitrary square matrix $\boldsymbol{A}$ can be written as the sum of a symmetric matrix $\boldsymbol{S}$ and a skew-symmetric matrix $\boldsymbol{S}^*$. It is easy to verify that $\boldsymbol{x}'\boldsymbol{S}^*\boldsymbol{x} = 0$. (Check!) It is therefore without loss of generality to consider quadratic forms $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ with a symmetric $\boldsymbol{A}$.

The quadratic form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ is said to be *positive definite* (*semi-definite*) if, and only if, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > (\geq)\ 0$ for all $\boldsymbol{x} \neq \boldsymbol{o}$. A square matrix $\boldsymbol{A}$ is said to be positive definite if its quadratic form is positive definite. Similarly, a matrix $\boldsymbol{A}$ is said to be *negative definite* (*semi-definite*) if, and only if, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} < (\leq)\ 0$ for all $\boldsymbol{x} \neq \boldsymbol{o}$. A matrix that is not definite or semi-definite is *indefinite*. A symmetric and positive semi-definite matrix is also known as a *Grammian matrix*. It can be shown that $\boldsymbol{A}$ is positive definite if, and only if, the principal minors,

$$\det(a_{11}), \qquad \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \qquad \det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \quad \cdots, \det(A),$$

are all positive; $\boldsymbol{A}$ is negative definite if, and only if, all the principal minors alternate in signs:

$$\det(a_{11}) < 0, \quad \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} > 0, \quad \det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} < 0, \quad \cdots.$$

Thus, a positive (negative) definite matrix must be nonsingular, but a positive determinant is *not* sufficient for a positive definite matrix. The difference between a positive (negative) definite matrix and a positive (negative) semi-definite matrix is that the latter may be singular. (Why?)

Theorem 5.3 *Let $\boldsymbol{A}$ be positive definite and $\boldsymbol{B}$ be nonsingular. Then $\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B}$ is also positive definite.*

Proof: For any $n \times 1$ matrix $\boldsymbol{y} \neq \boldsymbol{o}$, there exists $\boldsymbol{x} \neq \boldsymbol{o}$ such that $\boldsymbol{B}^{-1}\boldsymbol{x} = \boldsymbol{y}$. Hence,

$$\boldsymbol{y}'\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{y} = \boldsymbol{x}'\boldsymbol{B}^{-1\prime}(\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B})\boldsymbol{B}^{-1}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0. \quad \square$$

It follows that if $\boldsymbol{A}$ is positive definite, $\boldsymbol{A}^{-1}$ exists and $\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A}'(\boldsymbol{A}')^{-1}$ is also positive definite. It can be shown that a symmetric matrix is positive definite if, and only if, it can be factored as $\boldsymbol{P}'\boldsymbol{P}$, where $\boldsymbol{P}$ is a nonsingular matrix. Let $\boldsymbol{A}$ be a symmetric and positive definite matrix so that $\boldsymbol{A} = \boldsymbol{P}'\boldsymbol{P}$ and $\boldsymbol{A}^{-1} = \boldsymbol{P}^{-1}\boldsymbol{P}^{-1\prime}$. For any vector $\boldsymbol{x}$ and $\boldsymbol{w}$, let $\boldsymbol{u} = \boldsymbol{P}\boldsymbol{x}$ and $\boldsymbol{v} = \boldsymbol{P}^{-1\prime}\boldsymbol{w}$. Then

$$\begin{aligned}
(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})(\boldsymbol{w}'\boldsymbol{A}^{-1}\boldsymbol{w}) &= (\boldsymbol{x}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{x})(\boldsymbol{w}'\boldsymbol{P}^{-1}\boldsymbol{P}^{-1\prime}\boldsymbol{w}) \\
&= (\boldsymbol{u}'\boldsymbol{u})(\boldsymbol{v}'\boldsymbol{v}) \\
&\geq (\boldsymbol{u}'\boldsymbol{v})^2 \\
&= (\boldsymbol{x}'\boldsymbol{w})^2.
\end{aligned}$$

This result can be viewed as a generalization of the Cauchy-Schwartz inequality:

$$(\boldsymbol{x}'\boldsymbol{w})^2 \leq (\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})(\boldsymbol{w}'\boldsymbol{A}^{-1}\boldsymbol{w}).$$

5.4 Differentiation Involving Vectors and Matrices

Let $\boldsymbol{x}$ be an $n \times 1$ matrix, $f(\boldsymbol{x})$ a real function, and $\boldsymbol{f}(\boldsymbol{x})$ a vector-valued function with elements $f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})$. Then

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix}, \qquad \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}.$$

Some particular examples are:

1. $f(\boldsymbol{x}) = \boldsymbol{a}'\boldsymbol{x}$: $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{a}$.

2. $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{A}$ an $n \times n$ symmetric matrix: As

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_{ij}x_ix_j,$$

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}.$$

3. $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{A}$ an $m \times n$ matrix: $\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}'$.

4. $f(\boldsymbol{X}) = \operatorname{trace}(\boldsymbol{X})$ with $\boldsymbol{X}$ an $n \times n$ matrix: $\nabla_{\boldsymbol{X}} f(\boldsymbol{X}) = \boldsymbol{I}_n$.

5. $f(\boldsymbol{X}) = \det(\boldsymbol{X})$ with $\boldsymbol{X}$ an $n \times n$ matrix: $\nabla_{\boldsymbol{X}} f(\boldsymbol{X}) = \det(\boldsymbol{X})\, \boldsymbol{X}^{-1\prime}$.

When $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{A}$ a symmetric matrix, the matrix of second order derivatives, also known as the *Hessian* matrix, is

$$\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x}) = \nabla_{\boldsymbol{x}}(\nabla_{\boldsymbol{x}} f(\boldsymbol{x})) = \nabla_{\boldsymbol{x}}(2\boldsymbol{A}\boldsymbol{x}) = 2\boldsymbol{A}.$$

Analogous to the standard optimization problem, a necessary condition for maximizing (minimizing) the quadratic form $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ is $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{o}$, and a sufficient condition for a maximum (minimum) is that the Hessian matrix is negative (positive) definite.

5.5 Idempotent and Nilpotent Matrices

A square matrix $\boldsymbol{A}$ is said to be *idempotent* if $\boldsymbol{A} = \boldsymbol{A}^2$. Let $\boldsymbol{B}$ and $\boldsymbol{C}$ be two $n \times k$ matrices with rank $k < n$. An idempotent matrix can be constructed as $\boldsymbol{B}(\boldsymbol{C}'\boldsymbol{B})^{-1}\boldsymbol{C}'$; in particular, $\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'$ and $\boldsymbol{I}$ are idempotent. We observe that if an idempotent matrix $\boldsymbol{A}$ is nonsingular, then

$$\boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^2\boldsymbol{A}^{-1} = \boldsymbol{A}(\boldsymbol{A}\boldsymbol{A}^{-1}) = \boldsymbol{A}.$$

That is, all idempotent matrices are singular, except the identity matrix $\boldsymbol{I}$. It is also easy to see that if $\boldsymbol{A}$ is idempotent, then so is $\boldsymbol{I} - \boldsymbol{A}$.

A square matrix $\boldsymbol{A}$ is said to be *nilpotent* of index $r > 1$ if $\boldsymbol{A}^r = \boldsymbol{o}$ but $\boldsymbol{A}^{r-1} \neq \boldsymbol{o}$. For example, a lower (upper) triangular matrix with all diagonal elements equal to zero is called a sub-diagonal (super-diagonal) matrix. The sub-diagonal (super-diagonal) matrix is nilpotent of some index r.

5.6 Orthogonal Matrix

A square matrix $\boldsymbol{A}$ is *orthogonal* if $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}' = I$, i.e., $\boldsymbol{A}^{-1} = \boldsymbol{A}'$. Clearly, when $\boldsymbol{A}$ is orthogonal, $\boldsymbol{a}_i'\boldsymbol{a}_j = 0$ for $i \neq j$ and $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$. That is, the column (row) vectors of an orthogonal matrix are orthonormal. For example, the matrices we learned in Sections 4.1 and 4.3:

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix},$$

are orthogonal matrices. Given two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ and their orthogonal transformations $\boldsymbol{A}\boldsymbol{u}$ and $\boldsymbol{A}\boldsymbol{v}$. It is easy to see that $\boldsymbol{u}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{v} = \boldsymbol{u}'\boldsymbol{v}$ and that $\|\boldsymbol{u}\| = \|A\boldsymbol{u}\|$. Hence, orthogonal transformations preserve inner products, norms, angles, and distances. Applying this result to data matrices, we know sample variances, covariances, and correlation coefficients are invariant with respect to orthogonal transformations.

Note that when $\boldsymbol{A}$ is an orthogonal matrix,

$$1 = \det(\boldsymbol{I}) = \det(\boldsymbol{A})\det(\boldsymbol{A}') = (\det(A))^2,$$

so that $\det(\boldsymbol{A}) = \pm 1$. Hence, if $\boldsymbol{A}$ is orthogonal, $\det(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}') = \det(\boldsymbol{B})$ for any square matrix $\boldsymbol{B}$. If $\boldsymbol{A}$ is orthogonal and $\boldsymbol{B}$ is idempotent, then $\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}'$ is also idempotent because

$$(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}')(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}') = \boldsymbol{A}\boldsymbol{B}\boldsymbol{B}\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}'.$$

That is, pre- and post-multiplying a matrix by orthogonal matrices $\boldsymbol{A}$ and $\boldsymbol{A}'$ preserves determinant and idempotency. Also, the product of orthogonal matrices is again an orthogonal matrix.

A special orthogonal matrix is the *permutation* matrix which is obtained by rearranging the rows or columns of an identity matrix. A vector's components are permuted if this vector is multiplied by a permutation matrix. For example,

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_2 \\ x_1 \\ x_4 \\ x_3 \end{bmatrix}.$$

If $\boldsymbol{A}$ is a permutation matrix, then so is $\boldsymbol{A}^{-1} = \boldsymbol{A}'$; if $\boldsymbol{A}$ and $\boldsymbol{B}$ are two permutation matrices, then $\boldsymbol{A}\boldsymbol{B}$ is also a permutation matrix.

5.7 Projection Matrix

Let $V = V_1 \oplus V_2$ be a vector space. From Corollary 2.7, we can write $\boldsymbol{y}$ in V as $\boldsymbol{y} = \boldsymbol{y}_1 + \boldsymbol{y}_2$, where $\boldsymbol{y}_1 \in V_1$ and $\boldsymbol{y}_2 \in V_2$. For a matrix $\boldsymbol{P}$, the transformation $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{y}_1$ is called the *projection* of $\boldsymbol{y}$ onto V_1 along V_2 if, and only if, $\boldsymbol{P}\boldsymbol{y}_1 = \boldsymbol{y}_1$. The matrix $\boldsymbol{P}$ is called a *projection matrix*. The projection is said to be an *orthogonal projection* if, and only if, V_1 and V_2 are orthogonal complements. Hence, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are orthogonal. In this case, $\boldsymbol{P}$ is called an *orthogonal projection matrix* which projects vectors orthogonally onto the subspace V_1.

Theorem 5.4 *$\boldsymbol{P}$ is a projection matrix if, and only if, $\boldsymbol{P}$ is idempotent.*

Proof: Let $\boldsymbol{y}$ be a non-zero vector. Given that $\boldsymbol{P}$ is a projection matrix,

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{y}_1 = \boldsymbol{P}\boldsymbol{y}_1 = \boldsymbol{P}^2\boldsymbol{y},$$

so that $(\boldsymbol{P} - \boldsymbol{P}^2)\boldsymbol{y} = \boldsymbol{o}$. As $\boldsymbol{y}$ is arbitrary, we have $\boldsymbol{P} = \boldsymbol{P}^2$, an idempotent matrix. Conversely, if $\boldsymbol{P} = \boldsymbol{P}^2$,

$$\boldsymbol{P}\boldsymbol{y}_1 = \boldsymbol{P}^2\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} = \boldsymbol{y}_1.$$

Hence, $\boldsymbol{P}$ is a projection matrix. □

Let $\boldsymbol{p}_i$ denote the i thcolumn of $\boldsymbol{P}$. By idempotency of $\boldsymbol{P}$, $\boldsymbol{P}\boldsymbol{p}_i = \boldsymbol{p}_i$, so that $\boldsymbol{p}_i$ must be in V_1, the space to which the projection is made.

Theorem 5.5 *A matrix $\boldsymbol{P}$ is an orthogonal projection matrix if, and only if, $\boldsymbol{P}$ is symmetric and idempotent.*

Proof: Note that $\boldsymbol{y}_2 = \boldsymbol{y} - \boldsymbol{y}_1 = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$. If $\boldsymbol{y}_1$ is the orthogonal projection of $\boldsymbol{y}$,

$$0 = \boldsymbol{y}_1'\boldsymbol{y}_2 = \boldsymbol{y}'\boldsymbol{P}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.$$

Hence, $\boldsymbol{P}'(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{o}$ so that $\boldsymbol{P}' = \boldsymbol{P}'\boldsymbol{P}$ and $\boldsymbol{P} = \boldsymbol{P}'\boldsymbol{P}$. This shows that $\boldsymbol{P}$ is symmetric. Idempotency follows from the proof of Theorem 5.4. Conversely, if $\boldsymbol{P}$ is symmetric and idempotent,

$$\boldsymbol{y}_1'\boldsymbol{y}_2 = \boldsymbol{y}'\boldsymbol{P}'(\boldsymbol{y} - \boldsymbol{y}_1) = \boldsymbol{y}'(\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}^2\boldsymbol{y}) = \boldsymbol{y}'(\boldsymbol{P} - \boldsymbol{P}^2)\boldsymbol{y} = 0.$$

This shows that the projection is orthogonal. □

It is readily verified that $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ is an orthogonal projection matrix, where $\boldsymbol{A}$ is an $n \times k$ matrix with full column rank. Moreover, we have the following result; see Exercise 5.7

Theorem 5.6 *Given an $n \times k$ matrix with full column rank, $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ orthogonally projects vectors in $\Re^n$ onto* $\text{span}(\boldsymbol{A})$.

Clearly, if $\boldsymbol{P}$ is an orthogonal projection matrix, then so is $\boldsymbol{I} - \boldsymbol{P}$, which orthogonally projects vectors in $\Re^n$ onto $\text{span}(\boldsymbol{A})^{\perp}$. While $\text{span}(\boldsymbol{A})$ is k-dimensional, $\text{span}(\boldsymbol{A})^{\perp}$ is $(n-k)$-dimensional by Theorem 2.6.

5.8 Partitioned Matrix

A matrix may be partitioned into sub-matrices. Operations for partitioned matrices are analogous to standard matrix operations. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times n$ and $m \times m$ matrices, respectively. The direct sum of $\boldsymbol{A}$ and $\boldsymbol{B}$ is defined to be the $(n+m) \times (n+m)$ *block-diagonal* matrix:

$$\boldsymbol{A} \oplus \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{o} \\ \boldsymbol{o} & \boldsymbol{B} \end{bmatrix};$$

this can be generalized to the direct sum of finitely many matrices. Clearly, the direct sum is associative but not commutative, i.e., $\boldsymbol{A} \oplus \boldsymbol{B} \neq \boldsymbol{B} \oplus \boldsymbol{A}$.

Consider the following partitioned matrix:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}.$$

When either $\boldsymbol{A}_{11}$ or $\boldsymbol{A}_{22}$ is nonsingular, we have

$$\det(\boldsymbol{A}) = \det(\boldsymbol{A}_{11})\det(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}),$$

$$\det(\boldsymbol{A}) = \det(\boldsymbol{A}_{22})\det(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}).$$

If $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are nonsingular, let $\boldsymbol{Q} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$ and $\boldsymbol{R} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$. The inverse of the partitioned matrix $\boldsymbol{A}$ can be computed as:

$$\begin{aligned} \boldsymbol{A}^{-1} &= \begin{bmatrix} \boldsymbol{Q}^{-1} & -\boldsymbol{Q}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{Q}^{-1} & \boldsymbol{A}_{22}^{-1} - \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{Q}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A}_{11}^{-1} - \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{R}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{R}^{-1} \\ -\boldsymbol{R}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{R}^{-1} \end{bmatrix}. \end{aligned}$$

In particular, if $\boldsymbol{A}$ is block-diagonal so that $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{21}$ are zero matrices, then $\boldsymbol{Q} = \boldsymbol{A}_{11}$, $\boldsymbol{R} = \boldsymbol{A}_{22}$, and

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11}^{-1} & 0 \\ 0 & \boldsymbol{A}_{22}^{-1} \end{bmatrix}.$$

That is, the inverse of a block-diagonal matrix can be obtained by taking inverses of each block.

5.9 Statistical Applications

Consider now the problem of explaining the behavior of the dependent variable $\boldsymbol{y}$ using k linearly independent, explanatory variables: $\boldsymbol{X} = [\boldsymbol{\ell}, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k]$. Suppose that these variables contain $n > k$ observations so that $\boldsymbol{X}$ is an $n \times k$ matrix with rank k. The least squares method is to compute a regression "hyperplane", $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}$, that best fits the data $(\boldsymbol{y}\ \boldsymbol{X})$. Write $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, where $\boldsymbol{e}$ is the vector of residuals. Let

$$f(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$

Our objective is to minimize $f(\boldsymbol{\beta})$, the sum of squared residuals. As $\boldsymbol{X}$ is of rank k, $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular. In view of Section 5.4, the first order condition is:

$$\nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = -2\boldsymbol{X}'\boldsymbol{y} + 2(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{o},$$

which yields the solution

$$\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

The matrix of the second order derivatives is $2(\boldsymbol{X}'\boldsymbol{X})$, a positive definite matrix. Hence, the solution $\boldsymbol{\beta}$ minimizes $f(\boldsymbol{\beta})$ and is referred to as the *ordinary least squares* estimator. Note that

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

and that $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is an orthogonal projection matrix. The fitted regression hyperplane is in fact the orthogonal projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$. It is also easy to see that

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y},$$

which is orthogonal to the column space of $\boldsymbol{X}$. This fitted hyperplane is the best approximation of $\boldsymbol{y}$ in terms of the Euclidean norm, based on the "information" of $\boldsymbol{X}$.

Exercises

5.1 Let $\boldsymbol{A}$ be an $n \times n$ skew-symmetric matrix and $\boldsymbol{x}$ be $n \times 1$. Prove that $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = 0$.

5.2 Show that a positive definite matrix cannot be singular.

5.3 Consider the quadratic form $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ such that $\boldsymbol{A}$ is not symmetric. Find $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$.

5.4 Let $\boldsymbol{X}$ be an $n \times k$ matrix with full column rank and $\boldsymbol{\Sigma}$ be an $n \times n$ symmetric, positive definite matrix. Show that $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}$ is a projection matrix but not an orthogonal projection matrix.

5.5 Let $\boldsymbol{\ell}$ be the vector of n ones. Show that $\boldsymbol{\ell}\boldsymbol{\ell}'/n$ is an orthogonal projection matrix.

5.6 Let S_1 and S_2 be two subspaces of V such that $S_2 \subseteq S_1$. Let $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ be two orthogonal projection matrices projecting vectors onto S_1 and S_2, respectively. Find $\boldsymbol{P}_1\boldsymbol{P}_2$ and $(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P}_1)$.

5.7 Prove Theorem 5.6.

5.8 Given a dependent variable $\boldsymbol{y}$ and an explanatory variable $\boldsymbol{\ell}$, the vector of n ones. Apply the method in Section 2.4 and the least squares method to find the orthogonal projection of $\boldsymbol{y}$ on $\boldsymbol{\ell}$. Compare your results.

6 Eigenvalue and Eigenvector

In many applications it is important to transform a large matrix to a matrix of a simpler structure that preserves important properties of the original matrix.

6.1 Eigenvalue and Eigenvector

Let $\boldsymbol{A}$ be an $n \times n$ matrix. A non-zero vector $\boldsymbol{p}$ is said to be an *eigenvector* of $\boldsymbol{A}$ corresponding to an *eigenvalue* λ if

$$\boldsymbol{A}\boldsymbol{p} = \lambda\boldsymbol{p}$$

for some scalar λ. Eigenvalues and eigenvectors are also known as *latent roots* and *latent vectors* (*characteristic values* and *characteristic vectors*). In $\Re^3$, it is clear that when $\boldsymbol{A}\boldsymbol{p} = \lambda\boldsymbol{p}$, multiplication of $\boldsymbol{p}$ by $\boldsymbol{A}$ dilates or contracts $\boldsymbol{p}$. Note also that eigenvalues of a matrix need not be distinct; an eigenvalue may be repeated with *multiplicity* k.

Write $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{p} = \boldsymbol{o}$. In the light of Section 4.3, this homogeneous system has a non-trivial solution if

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = \boldsymbol{o}.$$

The equation above is known as the *characteristic polynomial* of $\boldsymbol{A}$, and its roots are eigenvalues. The solution space of $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{p} = \boldsymbol{o}$ characteristic polynomial is the *eigenspace* of $\boldsymbol{A}$ corresponding to λ. By Corollary 4.3, the dimension of the eigenspace is $n - \operatorname{rank}(\boldsymbol{A} - \lambda\boldsymbol{I})$. Note that if $\boldsymbol{p}$ is an eigenvector of $\boldsymbol{A}$, then so is $c\boldsymbol{p}$ for any non-zero scalar c. Hence, it is typical to normalize eigenvectors to unit length.

As an example, consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 3 & -2 & 0 \\ -2 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}.$$

It is easy to verify that $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = -(\lambda - 1)(\lambda - 5)^2 = 0$. The eigenvalues are thus 1 and 5 (with multiplicity 2). When $\lambda = 1$, we have

$$\begin{bmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \boldsymbol{o}.$$

It follows that $p_1 = p_2 = a$ for any a and $p_3 = 0$. Thus, $\{(1,1,0)'\}$ is a basis of the eigenspace corresponding to $\lambda = 1$. Similarly, when $\lambda = 5$,

$$\begin{bmatrix} -2 & -2 & 0 \\ -2 & -2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \boldsymbol{o}.$$

We have $p_1 = -p_2 = a$ and $p_3 = b$ for any a, b. Hence, $\{(1,-1,0)', (0,0,1)'\}$ is a basis of the eigenspace corresponding to $\lambda = 5$.

6.2 Diagonalization

Two $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be *similar* if there exists a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{B} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$, or equivalently $\boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^{-1} = \boldsymbol{A}$. The following results show that similarity transformation preserves many important properties of a matrix.

Theorem 6.1 *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two similar matrices. Then,*

(a) $\det(\boldsymbol{A}) = \det(\boldsymbol{B})$.

(b) $\operatorname{trace}(\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{B})$.

(c) $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same eigenvalues.

(d) $\boldsymbol{P}\boldsymbol{q}_B = \boldsymbol{q}_A$, where $\boldsymbol{q}_A$ and $\boldsymbol{q}_B$ are eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively.

(e) If $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular, then $\boldsymbol{A}^{-1}$ is similar to $\boldsymbol{B}^{-1}$.

Proof: Part (a), (b) and (e) are obvious. Part (c) follows because

$$\begin{aligned} \det(\boldsymbol{B} - \lambda \boldsymbol{I}) &= \det(\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} - \lambda \boldsymbol{P}^{-1}\boldsymbol{P}) \\ &= \det(\boldsymbol{P}^{-1}(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{P}) \\ &= \det(\boldsymbol{A} - \lambda \boldsymbol{I}). \end{aligned}$$

For part (d), we note that $\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{B}$. Then, $\boldsymbol{A}\boldsymbol{P}\boldsymbol{q}_B = \boldsymbol{P}\boldsymbol{B}\boldsymbol{q}_B = \lambda \boldsymbol{P}\boldsymbol{q}_B$. This shows that $\boldsymbol{P}\boldsymbol{q}_B$ is an eigenvector of $\boldsymbol{A}$. □

Of particular interest to us is the similarity between a square matrix and a diagonal matrix. A square matrix $\boldsymbol{A}$ is said to be *diagonalizable* if $\boldsymbol{A}$ is similar to a diagonal matrix $\boldsymbol{\Lambda}$, i.e., $\boldsymbol{\Lambda} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$ or equivalently $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$ for some nonsingular matrix $\boldsymbol{P}$. We also say that $\boldsymbol{P}$ *diagonalizes* $\boldsymbol{A}$. When $\boldsymbol{A}$ is diagonalizable, we have

$$\boldsymbol{A}\boldsymbol{p}_i = \lambda_i \boldsymbol{p}_i,$$

where $\boldsymbol{p}_i$ is the i thcolumn of $\boldsymbol{P}$ and λ_i is the i thdiagonal element of $\boldsymbol{\Lambda}$. That is, $\boldsymbol{p}_i$ is an eigenvector of $\boldsymbol{A}$ corresponding to the eigenvalue λ_i.

When $\boldsymbol{\Lambda} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$, these eigenvectors must be linearly independent. Conversely, if $\boldsymbol{A}$ has n linearly independent eigenvectors $\boldsymbol{p}_i$ corresponding to eigenvalues λ_i, $i = 1, \ldots, n$, we can write $\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda}$, where $\boldsymbol{p}_i$ is the i thcolumn of $\boldsymbol{P}$, and $\boldsymbol{\Lambda}$ contains diagonal terms λ_i. That $\boldsymbol{p}_i$ are linearly independent implies that $\boldsymbol{P}$ is invertible. It follows that $\boldsymbol{P}$ diagonalizes $\boldsymbol{A}$. We have proved:

Theorem 6.2 *Let $\boldsymbol{A}$ be an $n \times n$ matrix. $\boldsymbol{A}$ is diagonalizable if, and only if, $\boldsymbol{A}$ has n linearly independent eigenvectors.*

The result below indicates that if $\boldsymbol{A}$ has n distinct eigenvalues, the associated eigenvectors are linearly independent.

Theorem 6.3 *If $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ are eigenvectors of $\boldsymbol{A}$ corresponding to distinct eigenvalues $\lambda_1, \ldots, \lambda_n$, then $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\}$ is a linearly independent set.*

Proof: Suppose that $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ are linearly dependent. Let $1 \le r < n$ be the largest integer such that $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_r$ are linearly independent. Hence,

$$c_1\boldsymbol{p}_1 + \cdots + c_r\boldsymbol{p}_r + c_{r+1}\boldsymbol{p}_{r+1} = \boldsymbol{o},$$

where $c_i \neq 0$ for some $1 \le i \le r+1$. Multiplying both sides by $\boldsymbol{A}$, we obtain

$$c_1\lambda_1\boldsymbol{p}_1 + \cdots + c_{r+1}\lambda_{r+1}\boldsymbol{p}_{r+1} = \boldsymbol{o},$$

and multiplying by λ_{r+1} we have

$$c_1\lambda_{r+1}\boldsymbol{p}_1 + \cdots + c_{r+1}\lambda_{r+1}\boldsymbol{p}_{r+1} = \boldsymbol{o}.$$

Their difference is

$$c_1(\lambda_1 - \lambda_{r+1})\boldsymbol{p}_1 + \cdots + c_r(\lambda_r - \lambda_{r+1})\boldsymbol{p}_r = \boldsymbol{o}.$$

As $\boldsymbol{p}_i$ are linearly independent and eigenvalues are distinct, $c_1, \ldots, c_r$ must be zero. Consequently, $c_{r+1}\boldsymbol{p}_{r+1} = 0$ so that $c_{r+1} = 0$. This contradicts the fact that $c_i \neq 0$ for some i. $\square$

For an $n \times n$ matrix $\boldsymbol{A}$, that $\boldsymbol{A}$ has n distinct eigenvalues is a sufficient (but not necessary) condition for diagonalizability by Theorem 6.2 and 6.3. When some eigenvalues are equal, however, not much can be asserted in general. When $\boldsymbol{A}$ has n distinct eigenvalues, $\boldsymbol{\Lambda} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$ by Theorem 6.2. It follows that

$$\begin{aligned} \det(\boldsymbol{A}) = \det(\boldsymbol{\Lambda}) &= \prod_{i=1}^{n} \lambda_i, \\ \operatorname{trace}(\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{\Lambda}) &= \sum_{i=1}^{n} \lambda_i. \end{aligned}$$

In this case, $\boldsymbol{A}$ is nonsingular if, and only if, its eigenvalues are all non-zero. If we know that $\boldsymbol{A}$ is nonsingular, then by Theorem 6.1(e), $\boldsymbol{A}^{-1}$ is similar to $\boldsymbol{\Lambda}^{-1}$, and $\boldsymbol{A}^{-1}$ has eigenvectors $\boldsymbol{p}_i$ corresponding to eigenvalues $1/\lambda_i$. It is also easy to verify that $\boldsymbol{\Lambda} + c\boldsymbol{I} = \boldsymbol{P}^{-1}(\boldsymbol{A} + c\boldsymbol{I})\boldsymbol{P}$ and $\boldsymbol{\Lambda}^k = \boldsymbol{P}^{-1}\boldsymbol{A}^k\boldsymbol{P}$.

Finally, we note that when $\boldsymbol{P}$ diagonalizes $\boldsymbol{A}$, the eigenvectors of $\boldsymbol{A}$ form a new basis. The n-dimensional vector $\boldsymbol{x}$ can then be expressed as

$$\boldsymbol{x} = \xi_1 \boldsymbol{p}_1 + \cdots + \xi_n \boldsymbol{p}_n = \boldsymbol{P}\boldsymbol{\xi},$$

where $\boldsymbol{\xi}$ is the new coordinate vector of $\boldsymbol{x}$. In view of Section 4.1, $\boldsymbol{P}$ is the transition matrix from the new basis to the original Cartesian basis, and $\boldsymbol{P}^{-1}$ is the transition matrix from the Cartesian basis to the new basis. Each eigenvector is therefore the coordinate vector of the Cartesian unit vector with respect to the new basis vectors. Let $\boldsymbol{A}$ denote the $n \times n$ matrix corresponding to a linear transformation with respect to the Cartesian basis. Also let $\boldsymbol{A}^*$ denote the matrix corresponding to the same transformation with respect to the new basis. Then,

$$\boldsymbol{A}^*\boldsymbol{\xi} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}\boldsymbol{\xi},$$

so that $\boldsymbol{A}^* = \boldsymbol{\Lambda}$. Similarly, $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{P}\boldsymbol{A}^*\boldsymbol{P}^{-1}\boldsymbol{x}$. Hence, when $\boldsymbol{A}$ is diagonalizable, the corresponding linear transformation is rather straightforward when $\boldsymbol{x}$ is expressed in terms of the new basis vector.

6.3 Orthogonal Diagonalization

A square matrix $\boldsymbol{A}$ is said to be *orthogonally diagonalizable* if there is an orthogonal matrix $\boldsymbol{P}$ that diagonalizes $\boldsymbol{A}$, i.e., $\boldsymbol{\Lambda} = \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}$. In the light of the proof of Theorem 6.2, we have the following result.

Theorem 6.4 *Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then $\boldsymbol{A}$ is orthogonally diagonalizable, if and only if, $\boldsymbol{A}$ has an orthonormal set of n eigenvectors.*

If $\boldsymbol{A}$ is orthogonally diagonalizable, then $\boldsymbol{\Lambda} = \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}$ so that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$ is a symmetric matrix. The converse is also true, but its proof is more difficult and hence omitted. We have:

Theorem 6.5 *A matrix $\boldsymbol{A}$ is orthogonally diagonalizable if, and only if, $\boldsymbol{A}$ is symmetric.*

Moreover, we note that a symmetric matrix $\boldsymbol{A}$ has only *real* eigenvalues and eigenvectors. If an eigenvalue λ of an $n \times n$ symmetric matrix $\boldsymbol{A}$ is repeated with multiplicity k, then in view of Theorem 6.4 and Theorem 6.5, there must exist exactly k orthogonal eigenvectors corresponding to λ. Hence, this eigenspace is k-dimensional. It follows that $\operatorname{rank}(\boldsymbol{A} - \lambda \boldsymbol{I}) = n - k$. This implies that when $\lambda = 0$ is repeated with multiplicity k, $\operatorname{rank}(\boldsymbol{A}) = n - k$. This proves:

Theorem 6.6 *The number of non-zero eigenvalues of a symmetric matrix $\boldsymbol{A}$ is equal to* $\operatorname{rank}(\boldsymbol{A})$.

When $\boldsymbol{A}$ is orthogonally diagonalizable, we note that

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}' = \sum_{i=1}^{n} \lambda_i \boldsymbol{p}_i \boldsymbol{p}_i',$$

where $\boldsymbol{p}_i$ is the i^{th}column of $\boldsymbol{P}$. This is known as the *spectral (canonical) decomposition* of $\boldsymbol{A}$ which applies to both singular and nonsingular symmetric matrices. It can be seen that $\boldsymbol{p}_i \boldsymbol{p}_i'$ is an orthogonal projection matrix which orthogonally projects vectors onto $\boldsymbol{p}_i$.

We also have the following results for some special matrices. Let $\boldsymbol{A}$ be an orthogonal matrix and $\boldsymbol{p}$ be its eigenvector corresponding to the eigenvalue λ. Observe that

$$\boldsymbol{p}'\boldsymbol{p} = \boldsymbol{p}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{p} = \lambda^2 \boldsymbol{p}'\boldsymbol{p}.$$

Thus, the eigenvalues of an orthogonal matrix must be ± 1. It can also be shown that the eigenvalues of a positive definite (semi-definite) matrix are positive (non-negative); see Exercises 6.1 and 6.2. If $\boldsymbol{A}$ is symmetric and idempotent, then for any $\boldsymbol{x} \neq 0$,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{x} \geq 0.$$

That is, a symmetric and idempotent matrix must be positive semi-definite and therefore must have non-negative eigenvalues. In fact, as $\boldsymbol{A}$ is orthogonally diagonalizable, we have

$$\boldsymbol{\Lambda} = \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P} = \boldsymbol{\Lambda}^2.$$

Consequently, the eigenvalues of $\boldsymbol{A}$ must be either 0 or 1.

6.4 Generalization

Two n-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ are said to be orthogonal in the metric $\boldsymbol{B}$ if $\boldsymbol{x}'\boldsymbol{B}\boldsymbol{z} = 0$, where $\boldsymbol{B}$ is $n \times n$. A matrix $\boldsymbol{P}$ is orthogonal in the metric $\boldsymbol{B}$ if $\boldsymbol{P}'\boldsymbol{B}\boldsymbol{P} = \boldsymbol{I}$, i.e., $\boldsymbol{p}_i'\boldsymbol{B}\boldsymbol{p}_i = 1$ and $\boldsymbol{p}_i'\boldsymbol{B}\boldsymbol{p}_j = 0$ for $i \neq j$. Consider now the generalized eigenvalue problem with respect to $\boldsymbol{B}$:

$$(\boldsymbol{A} - \lambda\boldsymbol{B})\boldsymbol{x} = \boldsymbol{o}.$$

Again, this system has non-trivial solution if $\det(\boldsymbol{A} - \lambda\boldsymbol{B}) = 0$. In this case, λ is an eigenvalue of $\boldsymbol{A}$ in the metric $\boldsymbol{B}$ and $\boldsymbol{x}$ is an eigenvector corresponding to λ.

Theorem 6.7 *Let $\boldsymbol{A}$ be a symmetric matrix and $\boldsymbol{B}$ be a symmetric, positive definite matrix. Then there exists a diagonal matrix $\boldsymbol{\Lambda}$ and a matrix $\boldsymbol{P}$ orthogonal in the metric $\boldsymbol{B}$ such that $\boldsymbol{\Lambda} = \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}$.*

Proof: As $\boldsymbol{B}$ is positive definite, there exists a nonsingular matrix $\boldsymbol{C}$ such that $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{C}'$. Consider the symmetric matrix $\boldsymbol{C}^{-1}\boldsymbol{A}(\boldsymbol{C}')^{-1}$. Then there exists an orthogonal matrix $\boldsymbol{R}$ such that

$$\boldsymbol{R}'(\boldsymbol{C}^{-1}\boldsymbol{A}(\boldsymbol{C}')^{-1})\boldsymbol{R} = \boldsymbol{\Lambda},$$

with $\boldsymbol{R}'\boldsymbol{R} = \boldsymbol{I}$. By letting $\boldsymbol{P} = \boldsymbol{C}^{-1\prime}\boldsymbol{R}$, we have $\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P} = \boldsymbol{\Lambda}$ and

$$\boldsymbol{P}'\boldsymbol{B}\boldsymbol{P} = \boldsymbol{P}'\boldsymbol{C}\boldsymbol{C}'\boldsymbol{P} = \boldsymbol{R}'\boldsymbol{R} = \boldsymbol{I}. \quad \square$$

It follows that $\boldsymbol{A} = \boldsymbol{P}^{-1\prime}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$ and $\boldsymbol{B} = \boldsymbol{P}^{-1\prime}\boldsymbol{P}^{-1}$. Thus, when $\boldsymbol{A}$ is symmetric and $\boldsymbol{B}$ is symmetric and positive definite, $\boldsymbol{A}$ can be diagonalized by a matrix $\boldsymbol{P}$ that is orthogonal in the metric $\boldsymbol{B}$. Note that $\boldsymbol{P}$ is not orthogonal in the usual sense, i.e., $\boldsymbol{P}' \neq \boldsymbol{P}^{-1}$. This result shows that for the eigenvalues of $\boldsymbol{A}$ in the metric $\boldsymbol{B}$, the associated eigenvectors are orthonormal in the metric $\boldsymbol{B}$.

6.5 Rayleigh Quotients

Let $\boldsymbol{A}$ be a symmetric matrix. The *Rayleigh quotient* is

$$q = \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}.$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ be the corresponding eigenvectors. Let $\boldsymbol{z} = \boldsymbol{P}'\boldsymbol{x}$. Then

$$\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \frac{\boldsymbol{x}'\boldsymbol{P}\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{P}\boldsymbol{P}'\boldsymbol{x}} = \frac{\boldsymbol{z}'\boldsymbol{\Lambda}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}},$$

so that the difference between the Rayleigh quotient and the largest eigenvalue is

$$q - \lambda_1 = \frac{\boldsymbol{z}'(\boldsymbol{\Lambda} - \lambda_1 \boldsymbol{I})\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}} = \frac{(\lambda_2 - \lambda_1)z_2^2 + \cdots + (\lambda_n - \lambda_1)z_n^2}{z_1^2 + z_2^2 + \cdots + z_n^2} \leq 0.$$

That is, the Rayleigh quotient $q \leq \lambda_1$. Similarly, we can show that $q \geq \lambda_n$. By setting $\boldsymbol{x} = \boldsymbol{p}_1$ and $\boldsymbol{x} = \boldsymbol{p}_n$, the resulting Rayleigh quotients are λ_1 and λ_n, respectively. This proves:

Theorem 6.8 *Let $\boldsymbol{A}$ be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then*

$$\lambda_n = \min_{\boldsymbol{x} \neq \boldsymbol{o}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}, \qquad \lambda_1 = \max_{\boldsymbol{x} \neq \boldsymbol{o}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}.$$

6.6 Vector and Matrix Norms

Let V be a vector space. The function $\| \cdot \| : V \to [0, \infty)$ is a *norm* on V if it satisfies the following properties:

1. $\|\boldsymbol{v}\| = 0$ if, and only if, $\boldsymbol{v} = \boldsymbol{o}$;
2. $\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$;
3. $\|h\boldsymbol{v}\| = |h|\|\boldsymbol{v}\|$, where h is a scalar.

We are already familiar with the Euclidean norm; there are many other norms. For example, for an n-dimensional vector $\boldsymbol{v}$, its ℓ_1, ℓ_2 (Euclidean), and ℓ_∞ (maximal) norms are:

$$\begin{aligned} \|\boldsymbol{v}\|_1 &= \textstyle\sum_{i=1}^n |v_i|; \\ \|\boldsymbol{v}\|_2 &= \left(\textstyle\sum_{i=1}^n v_i^2\right)^{1/2}; \\ \|\boldsymbol{v}\|_\infty &= \max_{1 \leq i \leq n} |v_i|. \end{aligned}$$

It is easy to verify that these norms satisfy the above properties.

A *matrix norm* of a matrix $\boldsymbol{A}$ is a non-negative number $\|\boldsymbol{A}\|$ such that

1. $\|\boldsymbol{A}\| = 0$ if, and only if, $\boldsymbol{A} = \boldsymbol{o}$.
2. $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$.
3. $\|h\boldsymbol{A}\| = |h|\|\boldsymbol{A}\|$, where h is a scalar.
4. $\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{B}\|$.

A matrix norm $\|\boldsymbol{A}\|$ is said to be compatible with a vector norm $\|\boldsymbol{v}\|$ if $\|\boldsymbol{A}\boldsymbol{v}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{v}\|$. Thus,

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{v}\neq\boldsymbol{o}} \frac{\|\boldsymbol{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$$

is known as a natural matrix norm associated with the vector norm. Let $\boldsymbol{A}$ be an $n \times n$ matrix. The ℓ_1, ℓ_2, Euclidean, and ℓ_∞ norms of $\boldsymbol{A}$ are:

$$\|\boldsymbol{A}\|_1 = \max_{1\leq j\leq n} \textstyle\sum_{i=1}^n |a_{ij}|;$$

$$\|\boldsymbol{A}\|_2 = (\lambda_1)^{1/2}, \text{ where } \lambda_1 \text{ is the largest eigenvalue of } \boldsymbol{A}'\boldsymbol{A};$$

$$\|\boldsymbol{A}\|_E = \text{trace}(\boldsymbol{A}'\boldsymbol{A})^{1/2};$$

$$\|\boldsymbol{A}\|_\infty = \max_{1\leq i\leq n} \textstyle\sum_{j=1}^n |a_{ij}|.$$

If $\boldsymbol{A}$ is an $n \times 1$ matrix, these matrix norms are just the corresponding vector norms defined earlier.

6.7 Statistical Applications

Let $\boldsymbol{x}$ be an n-dimensional normal random vector with mean zero and covariance matrix $\boldsymbol{I}_n$. It is well known that $\boldsymbol{x}'\boldsymbol{x} = \sum_{i=1}^n x_i^2$ is a χ^2 random variable with n degrees of freedom. If $\boldsymbol{A}$ is a symmetric and idempotent matrix with rank r, then it can be diagonalized by an orthogonal matrix $\boldsymbol{P}$ such that

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{P}(\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P})\boldsymbol{P}'\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'\boldsymbol{x},$$

where the diagonal elements of $\boldsymbol{\Lambda}$ are either zero or one. As orthogonal transformations preserve rank, $\boldsymbol{\Lambda}$ must have r eigenvalues equal to one. Let $\boldsymbol{P}'\boldsymbol{x} = \boldsymbol{y}$. Then, $\boldsymbol{y}$ is also normally distributed with mean zero and covariance matrix $\boldsymbol{P}'\boldsymbol{I}_n\boldsymbol{P} = \boldsymbol{I}_n$. Without loss of generality we can write

$$\boldsymbol{x}'\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'\boldsymbol{x} = \boldsymbol{y}'\begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{o} \\ \boldsymbol{o} & \boldsymbol{o} \end{bmatrix}\boldsymbol{y} = \sum_{i=1}^r y_i^2,$$

which is clearly a χ^2 random variable with r degrees of freedom.

Exercises

6.1 Show that a matrix is positive definite if, and only if, its eigenvalues are all positive.

6.2 Show that a matrix is positive semi-definite but not positive definite if, and only if, at least one of its eigenvalue is zero while the remaining eigenvalues are positive.

6.3 Let $\boldsymbol{A}$ be a symmetric and idempotent matrix. Show that trace($\boldsymbol{A}$) is the number of non-zero eigenvalues of $\boldsymbol{A}$ and rank($\boldsymbol{A}$) = trace($\boldsymbol{A}$).

6.4 Let $\boldsymbol{P}$ be the orthogonal matrix such that $\boldsymbol{P}'(\boldsymbol{A}'\boldsymbol{A})\boldsymbol{P} = \boldsymbol{\Lambda}$, where $\boldsymbol{A}$ is $n \times k$ with rank $k < n$. What are the properties of $\boldsymbol{Z}^* = \boldsymbol{A}\boldsymbol{P}$ and $\boldsymbol{Z} = \boldsymbol{Z}^*\boldsymbol{\Lambda}^{-1/2}$? Note that the column vectors of $\boldsymbol{Z}^*$ ($\boldsymbol{Z}$) are known as the (standardized) *principal axes* of $\boldsymbol{A}'\boldsymbol{A}$.

6.5 Given the information in the question above, show that the non-zero eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}'$ are equal. Also show that the eigenvectors associated with the non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}'$ are the standardized principal axes of $\boldsymbol{A}'\boldsymbol{A}$.

6.6 Given the information in the question above, show that

$$\boldsymbol{A}\boldsymbol{A}' = \sum_{i=1}^{k} \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i',$$

where $\boldsymbol{z}_i$ is the i thcolumn of $\boldsymbol{Z}$.

6.7 Given the information in the question above, show that the eigenvectors associated with the non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}'$ and $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ are equal.

6.8 Given the information in the question above, show that

$$\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}' = \sum_{i=1}^{k} \boldsymbol{z}_i \boldsymbol{z}_i',$$

where $\boldsymbol{z}_i$ is the i thcolumn of $\boldsymbol{Z}$.

Answers to Exercises

Chapter 1

1. $\boldsymbol{u} + \boldsymbol{v} = (5, -1, 3)$, $\boldsymbol{u} - \boldsymbol{v} = (-3, -5, 1)$.

2. Let $\boldsymbol{u} = (-3, 2, 4)$, $\boldsymbol{v} = (-3, 0, 0)$ and $\boldsymbol{w} = (0, 0, 4)$. Then $\|\boldsymbol{u}\| = \sqrt{29}$, $\|\boldsymbol{v}\| = 3$, $\|\boldsymbol{w}\| = 4$, $\boldsymbol{u} \cdot \boldsymbol{v} = 9$, $\boldsymbol{u} \cdot \boldsymbol{w} = 16$, and $\boldsymbol{v} \cdot \boldsymbol{w} = 0$.

3. $\theta = \cos^{-1}\left(\frac{13}{2\sqrt{77}}\right)$.

4. $(2/\sqrt{13}, 3/\sqrt{13})$, $(-2/\sqrt{13}, -3/\sqrt{13})$.

5. By the fourth property of inner products, $\|\boldsymbol{u}\| = (\boldsymbol{u} \cdot \boldsymbol{u})^{1/2} = 0$ if, and only if, $\boldsymbol{u} = \boldsymbol{o}$.

6. By the triangle inequality,
$$\|\boldsymbol{u}\| = \|\boldsymbol{u} - \boldsymbol{v} + \boldsymbol{v}\| \leq \|\boldsymbol{u} - \boldsymbol{v}\| + \|\boldsymbol{v}\|,$$
$$\|\boldsymbol{v}\| = \|\boldsymbol{v} - \boldsymbol{u} + \boldsymbol{u}\| \leq \|\boldsymbol{v} - \boldsymbol{u}\| + \|\boldsymbol{u}\| = \|\boldsymbol{u} - \boldsymbol{v}\| + \|\boldsymbol{u}\|.$$
The inequality holds as an equality when $\boldsymbol{u}$ and $\boldsymbol{v}$ are linearly dependent.

7. Apply the triangle inequality repeatedly.

8. The Cauchy-Schwartz inequality gives $|s_{x,y}| \leq s_x s_y$; the triangle inequality yields $s_{x+y} \leq s_x + s_y$.

Chapter 2

1. As a vector space is closed under vector addition, hence for $\boldsymbol{u} \in V$, $\boldsymbol{u} + (-\boldsymbol{u}) = \boldsymbol{o}$ must be in V.

2. Both (a) and (b) are subspaces of $\mathbb{R}^3$ because they are closed under vector addition and scalar multiplication.

3. Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$, $r < k$, be a set of linearly dependent vectors. Hence, there exist $a_1, \ldots, a_r$, which are not all zeros, such that
$$a_1\boldsymbol{u}_1 + \cdots + a_r\boldsymbol{u}_r = \boldsymbol{o}.$$
Thus, $a_1, \ldots, a_r$ together with $k - r$ zeros form a solution to
$$c_1\boldsymbol{u}_1 + \cdots + c_r\boldsymbol{u}_r + c_{r+1}\boldsymbol{u}_{r+1} + \cdots + c_k\boldsymbol{u}_k = \boldsymbol{o}.$$

Hence, S is linearly dependent, proving (a). For part (b), suppose $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$, $r < k$, is an arbitrary set of linearly dependent vectors. Then S is linearly dependent by (a), contradicting the original hypothesis.

4. Let $S = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ be a basis for V and $Q = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ a set in V with $m > n$. We can write $\boldsymbol{v}_i = a_{1i}\boldsymbol{u}_1 + \cdots + a_{ni}\boldsymbol{u}_n$ for all $i = 1, \ldots, m$. Hence,

$$\begin{aligned}
\mathbf{0} &= c_1\boldsymbol{v}_1 + \cdots + c_m\boldsymbol{v}_m \\
&= c_1(a_{11}\boldsymbol{u}_1 + \cdots + a_{n1}\boldsymbol{u}_n) + c_2(a_{12}\boldsymbol{u}_1 + \cdots + a_{n2}\boldsymbol{u}_n) + \cdots + \\
&\qquad\quad c_m(a_{1m}\boldsymbol{u}_1 + \cdots + a_{nm}\boldsymbol{u}_n) \\
&= (c_1a_{11} + c_2a_{12} + \cdots + c_ma_{1m})\boldsymbol{u}_1 + \cdots + \\
&\qquad\quad (c_1a_{n1} + c_2a_{n2} + \cdots + c_ma_{nm})\boldsymbol{u}_n.
\end{aligned}$$

As S is a basis, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are linearly dependent so that

$$\begin{aligned}
c_1a_{11} + c_2a_{12} + \cdots + c_ma_{1m} &= 0 \\
c_1a_{21} + c_2a_{22} + \cdots + c_ma_{2m} &= 0 \\
\vdots \quad &\quad \vdots \\
c_1a_{n1} + c_2a_{n2} + \cdots + c_ma_{nm} &= 0.
\end{aligned}$$

This is a system of n equations with m unknowns in c. As $m > n$, this system has infinitely many solutions. Thus, $c_i \neq 0$ for some i, and Q is linearly dependent.

5. (a) is a basis because $a_1(2,1) + a_2(3,0) = 0$ implies that $a_1 = a_2 = 0$; (b) is not a basis because there are infinitely many solutions for $a_1(3,9) + a_2(-4,-12) = 0$, e.g., $a_1 = 4$ and $a_2 = 3$.

6. $\boldsymbol{w} = (-4/5, 2, 3/5)$, $\boldsymbol{e} = (9/5, 0, 12/5)$.

7. $\boldsymbol{v}_1 = (0, 2, 1, 0)$, $\boldsymbol{v}_2 = (1, -1/5, 2/5, 0)$, $\boldsymbol{v}_3 = (1/2, 1/2, -1, -1)$, $\boldsymbol{v}_3 = (4/15, 4/15, -8/15, 4/5)$.

8. The regression line is $\hat{\boldsymbol{y}} = \beta\boldsymbol{x}$ with $\beta = \boldsymbol{y} \cdot \boldsymbol{x}/\boldsymbol{x} \cdot \boldsymbol{x}$.

Chapter 3

1. It suffices to note that $(\boldsymbol{AB})_{11} = 53$ and $(\boldsymbol{BA})_{11} = 6$. Hence, $\boldsymbol{AB} \neq \boldsymbol{BA}$.

2. For example,

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 4 & 6 \\ -2 & -3 \end{bmatrix},$$

and $\boldsymbol{AB} = 0$. Note that both matrices are singular.

3. For example,

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} 4 \\ -2 \end{bmatrix}, \qquad \boldsymbol{C} = \begin{bmatrix} 6 \\ -3 \end{bmatrix}.$$

Clearly, $\boldsymbol{AB} = \boldsymbol{AC} = \boldsymbol{o}$, but $\boldsymbol{B} \neq \boldsymbol{C}$.

4. The cofactors along the first row are

$$\begin{aligned} C_{11} &= a_{22}a_{33} - a_{23}a_{32}, \qquad C_{12} = -(a_{21}a_{33} - a_{23}a_{31}), \\ C_{13} &= a_{21}a_{32} - a_{22}a_{31}. \end{aligned}$$

Thus,

$$\begin{aligned} \det(\boldsymbol{A}) &= a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} \\ &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\quad - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}. \end{aligned}$$

5. $\text{trace}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{trace}(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = \text{trace}(\boldsymbol{I}_k) = k$.

6. Apply the inequalities of Section 3.4 repeatedly, we have $\text{rank}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = k$ and $\text{rank}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = k$.

7. As $\boldsymbol{A}$ is symmetric, $\boldsymbol{A} = \boldsymbol{A}'$ so that $\boldsymbol{A}^{-1} = (\boldsymbol{A}')^{-1} = (\boldsymbol{A}^{-1})'$ is symmetric. It follows that $\text{adj}(\boldsymbol{A})$ is also symmetric because $\boldsymbol{A}^{-1} = \text{adj}(\boldsymbol{A})/\det(\boldsymbol{A})$. Alternatively, by evaluating $\text{adj}(\boldsymbol{A})$ directly one can find that the adjoint matrix is symmetric.

8. The adjoint matrix of $\boldsymbol{A}$ is

$$\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

Hence,

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

Chapter 4

1. The first three vectors are all in range(T); the vector $(3, -8, 2, 0)$ is in ker(T).

2. (i) $\text{rank}(T) = n$ and $\text{nullity}(T) = 0$.
   (ii) $\text{rank}(T) = 0$ and $\text{nullity}(T) = n$.
   (iii) $\text{rank}(T) = n$ and $\text{nullity}(T) = 0$.

3. (i) $(x, y) \to (2x, y)$ is an expansion in the x-direction with factor 2.
   (ii) $(x, y) \to (x, y/2)$ is a compression in the y-direction with factor 1/2.
   (iii) $(x, y) \to (x + 2y, y)$ is a shear in the x-direction with factor 2.
   (iv) $(x, y) \to (x, x/2 + y)$ is a shear in the y-direction with factor 1/2.

4. $\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{u} = (5, 2)$; $\boldsymbol{A}_2\boldsymbol{A}_1\boldsymbol{u} = (1, 4)$. These transformations involve a shear in the x-direction and a reflection about the line $x = y$; their order does matter.

Chapter 5

1. The elements of a skew-symmetric matrix $\boldsymbol{A}$ are such that $a_{ii} = 0$ for all i and $a_{ij} = -a_{ji}$ for all $i \neq j$. Hence,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(a_{ij} + a_{ji})x_i x_j = 0.$$

2. If $\boldsymbol{A}$ is singular, then there exists a $\boldsymbol{x} \neq \boldsymbol{o}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{o}$. For this particular $\boldsymbol{x}$, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = 0$. Hence, $\boldsymbol{A}$ is not positive definite. Note, however, that this does not violate the condition for positive semi-definiteness.

3. As $f(\boldsymbol{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$,

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \sum_{j=1}^{n}(a_{1j} + a_{j1})x_j \\ \sum_{j=1}^{n}(a_{2j} + a_{j2})x_j \\ \vdots \\ \sum_{j=1}^{n}(a_{nj} + a_{jn})x_j \end{bmatrix} = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}.$$

4. Clearly, $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}$ is idempotent but not symmetric.

5. It is easy to see that $\mathbf{1}\mathbf{1}'/n$ is symmetric and idempotent.

6. AS $\boldsymbol{P}_2 \in S_2 \subseteq S_1$, $\boldsymbol{P}_1\boldsymbol{P}_2 = \boldsymbol{P}_2$. Similarly, as $(\boldsymbol{I} - \boldsymbol{P}_1) \in S_1^{\perp} \subseteq S_2^{\perp}$, we have $(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P}_1) = \boldsymbol{I} - \boldsymbol{P}_1$.

7. For any vector $\boldsymbol{y} \in \text{span}(\boldsymbol{A})$, we can write $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{c}$ for some non-zero vector $\boldsymbol{c}$. Then, $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{y} = \boldsymbol{A}\boldsymbol{c} = \boldsymbol{y}$; this shows that this transformation is indeed a projection. For $\boldsymbol{y} \in \text{span}(\boldsymbol{A})^{\perp}$, $\boldsymbol{A}'\boldsymbol{y} = \boldsymbol{o}$ so that $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{y} = \boldsymbol{o}$. Hence, the projection must be orthogonal.

8. As we have only one explanatory variable $\mathbf{1}$, the orthogonal projection matrix is $\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \mathbf{1}\mathbf{1}'/n$, and the orthogonal projection of $\boldsymbol{y}$ on $\mathbf{1}$ is $(\mathbf{1}\mathbf{1}'/n)\boldsymbol{y} = \mathbf{1}\bar{\boldsymbol{y}}$.

Chapter 6

1. Consider the quadratic form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$. We can take $\boldsymbol{A}$ as a symmetric matrix and let $\boldsymbol{P}$ orthogonally diagonalize $\boldsymbol{A}$. For any $\boldsymbol{x} \neq \boldsymbol{o}$, we can find a $\boldsymbol{y}$ such that $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y}$ and

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}'\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{i=1}^{n} \lambda_i y_i^2.$$

   Clearly, if $\lambda_i > 0$ for all i, then $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$. Conversely, suppose $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{o}$. Let $\boldsymbol{e}_i$ denote the i thCartesian unit vector. Then for $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{e}_i$, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ implies $\lambda_i > 0$, $i = 1, \ldots, n$.

2. Without loss of generality assume that $\lambda_1 = 0$ and that other eigenvalues are positive. Then in view of the above proof, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$, we have $\lambda_i \geq 0$. If there is no $\lambda_i = 0$, then $\boldsymbol{A}$ must be positive definite, contradicting the original hypothesis. Hence, there must be at least one $\lambda_i = 0$.

3. As the eigenvalues of $\boldsymbol{A}$ are either zero or one, trace($\boldsymbol{\Lambda}$) is just the number of non-zero eigenvalues (say r). We also know that similarity transformations preserve rank and trace. It follows that rank($\boldsymbol{A}$) = rank($\boldsymbol{\Lambda}$) = r and trace($\boldsymbol{A}$) = trace($\boldsymbol{\Lambda}$) = r.

4. Note that $\boldsymbol{Z}^*$ and $\boldsymbol{Z}$ are $n \times k$ and that their column vectors are linear combinations of the column vectors of $\boldsymbol{A}$. Let $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1/2}$. Then $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{B}^{-1}$, so that each column vector of $\boldsymbol{A}$ is a linear combination of the column vectors of $\boldsymbol{Z}$. As $\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^* = \boldsymbol{\Lambda}$, $\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{I}_k$. Hence, the column vectors of $\boldsymbol{Z}$ form an orthonormal basis of span($\boldsymbol{A}$).

5. As rank($\boldsymbol{A}\boldsymbol{A}'$) = k, then by Theorem 6.6, $\boldsymbol{A}\boldsymbol{A}'$ has only k non-zero eigenvalues. Given $(\boldsymbol{A}'\boldsymbol{A})\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda}$, we can premultiply and postmultiply this expression by $\boldsymbol{A}$ and $\boldsymbol{\Lambda}^{-1/2}$, respectively, and obtain $(\boldsymbol{A}\boldsymbol{A}')\boldsymbol{A}\boldsymbol{P}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}$. That is,

$(\boldsymbol{A}\boldsymbol{A}')\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{\Lambda}$, where $\boldsymbol{Z}$ contains k orthonormal eigenvectors of $\boldsymbol{A}\boldsymbol{A}'$, and $\boldsymbol{\Lambda}$ is the matrix of non-zero eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}'$.

6. By orthogonal diagonalization, $(\boldsymbol{A}\boldsymbol{A}')\boldsymbol{C} = \boldsymbol{C}\boldsymbol{D}$, where

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{Z}^+ \end{bmatrix}, \boldsymbol{D} = \begin{bmatrix} \boldsymbol{\Lambda} & \boldsymbol{o} \\ \boldsymbol{o} & \boldsymbol{o} \end{bmatrix},$$

with $\boldsymbol{Z}^+$ the matrix of eigenvectors associated with the zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}'$. Hence,

$$\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}' = \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}' = \sum_{i=1}^{k} \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i'.$$

7. It can be seen that

$$\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}'\boldsymbol{A}'[\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}']\boldsymbol{A}\boldsymbol{P}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}'(\boldsymbol{A}'\boldsymbol{A})\boldsymbol{P}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{I}_k.$$

Hence, $\boldsymbol{Z}$ are also eigenvectors $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$, corresponding to the eigenvalues equal to one.

8. As in previous proofs,

$$\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}' = \boldsymbol{Z}\boldsymbol{Z}' = \sum_{i=1}^{k} \boldsymbol{z}_i \boldsymbol{z}_i'.$$

Note that $\boldsymbol{z}_i\boldsymbol{z}_i'$ orthogonally projects vectors onto $\boldsymbol{z}_i$. Hence, $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ orthogonally projects vectors onto span$(\boldsymbol{A})$.

Bibliography

Cited References

Anton, Howard (1981). *Elementary Linear Algebra*, third ed., New York: Wiley.

Basilevsky, Alexander (1983). *Applied Matrix Algebra in the Statistical Sciences*, New York: North-Holland.

Other Readings

Graybill, Franklin A. (1969). *Matrix with Applications in Statistics*, Belmont: Wadswoth.

Magnus, Jan R. & Heinz Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Chichester: Wiley.

Noble, Ben & James W. Daniel (1977). *Applied Matrix Algebra*, second ed., Englewood Cliffs: Prentice-Hall.

Pollard, D. S. G. (1979). *The Algebra of Econometrics*, Chichester: Wiley.