ECOM40006/ECOM90013 Econometrics 3
Department of Economics
University of Melbourne

An Introduction to Matrices

Semester 1, 2025

Contents

Preface v

1 Matrix Algebra 1
- 1.1 Definition and Notation . . . 1
- 1.2 Special Cases . . . 2
- 1.3 Matrix Operations . . . 4
  - 1.3.1 Matrix Transpose . . . 5
  - 1.3.2 Scalar Multiplication . . . 5
  - 1.3.3 Matrix Addition . . . 6
  - 1.3.4 Matrix Equality . . . 6
  - 1.3.5 Matrix Multiplication . . . 7
  - 1.3.6 Trace of a (Square) Matrix . . . 10
  - 1.3.7 Matrix Inverse . . . 10
- 1.4 Matrices With Special Properties Under Various Matrix Operations 12
  - 1.4.1 Idempotents . . . 12
  - 1.4.2 Orthogonal and Orthonormal Matrices . . . 14
- 1.5 Linearly Independent Vectors . . . 16
- 1.6 An Introduction to Solving Matrix Equations . . . 19

2 Elementary Matrix Operations and Matrix Reduction 25
- 2.1 Reduction of Matrices . . . 25
- 2.2 Elementary Matrix Operations . . . 25
- 2.3 Solving Systems by Reduction When the Solution is Unique . . . 32
- 2.4 Solving Systems by Reduction: The General Case . . . 34
  - 2.4.1 Introduction . . . 34
  - 2.4.2 More Equations Than Variables . . . 34
  - 2.4.3 Fewer Equations Than Variables . . . 35
  - 2.4.4 A Singular Coefficient Matrix . . . 36
- 2.5 Rank of a Matrix . . . 37
  - 2.5.1 Definition and Notation . . . 37
  - 2.5.2 Characterizing Solutions to Systems of Equations in Terms of Rank . . . 38

3 Determinants and Cramer's Rule 39
- 3.1 Determinants of 2×2 and 3×3 Matrices . . . 39
- 3.2 Properties of Determinants . . . 41
- 3.3 The Laplace Expansion of a Determinant . . . 43
- 3.4 The Diagonal Expansion of the Determinant of a Matrix . . . 45

3.5 Matrix Inversion and Determinants . . . 48
3.6 Cramer's Rule . . . 49
3.6.1 A Formal Development of Cramer's Rule . . . 51
3.7 Quadratic Forms and Definiteness . . . 52
3.8 The Determinant of a Vandermonde Matrix . . . 53

4 Partitioned Matrices 56
4.1 Definition and Elementary Operations . . . 56
4.2 Determinant of a Partitioned Matrix . . . 57
4.3 Partitioned Inverse . . . 58
4.4 Schur Complements . . . 61
4.5 Some Properties of Triangular Matrices . . . 61
4.6 More On Solving Systems of Linear Equations . . . 62
4.7 The Frisch-Waugh-Lovell Theorem . . . 65

5 The Generalized Inverse 70
5.1 The Generalized Inverse In General . . . 70
5.2 The Moore-Penrose Generalized Inverse . . . 73
5.3 The Left and Right Inverse . . . 73
5.4 Solving Linear Systems of Equations Using a Generalized Inverse . . . 74
5.5 Generalized Inverses and Least Squares . . . 75

6 Eigenvalues and Eigenvectors 80
6.1 That Which We Call a Rose By Any Other Word Would Smell as Sweet . . . 80
6.2 Motivation . . . 80
6.3 The Characteristic Equation of $\mathbf{A}$ . . . 81
6.4 Eigenfunctions of Symmetric Matrices . . . 86
6.5 The Elementary Symmetric Functions . . . 87

7 Some Useful Matrix Decompositions 89
7.1 Equivalent Canonical Form . . . 89
7.2 Full Rank Factorization . . . 90
7.3 Symmetric Matrices . . . 92
7.3.1 Congruent Canonical Form of a Symmetric Matrix . . . 92
7.3.2 The Spectral Decomposition of a Symmetric Matrix . . . 95

8 Kronecker Products and Vectorization 97
8.1 Kronecker Product . . . 97
8.2 Vectorization: vec and vech . . . 98

9 Matrix Calculus: Derivatives and Differentials 101
9.1 Some Fundamentals and Notation . . . 101
9.2 Notational Conventions . . . 103
9.3 Some Well-Known Results . . . 104
9.4 Standard Rules of Differentiation . . . 110
9.4.1 Scalar-Valued Functions of Matrix Argument . . . 111
9.5 Differentials . . . 111
9.6 Jacobians of Transformation . . . 114

9.7 Jacobians of Transformation and Exterior Differential Forms 116

10 Input-Output Matrices **120**

Bibliography **123**

Preface

Matrix algebra is concerned with the manipulation of collections of numbers. Perhaps the most attractive feature of matrices, for our purposes, is that matrix operations provide an extremely economical notation for certain calculations that are of interest to us. In particular, systems of linear equations (which may include the first order conditions from multivariate optimization problems) are often most easily represented and manipulated in matrix notation. The rules of matrix algebra are very similar to those for real numbers, although there are differences that can be important. Nevertheless, most of the time, treating matrices in exactly the same way that you treat numbers will bring you to the correct answer.

These Notes have grown to meet the needs of a number of different subjects and so nobody should expect to need to read it all. Nor should people attempt to remember all that is in here. They are provided as a resource that you can dip into as required. Moreover, they are far from complete and continue to evolve. The aim is not to replace any of the existing texts on matrix algebra, but simply to gather some important results together in a somewhat smaller form. That said, it would be remiss of me not to mention my heavy reliance on Searle (1982), which I have always found to be an incredibly useful text to have close at hand.

Chris Skeels
Melbourne
March 12, 2025

Chapter 1

Matrix Algebra

1.1 Definition and Notation

Definition 1.1.1. Matrix.
A *matrix* of order $m \times n$ is a rectangular array consisting of m rows and n columns, where m and n are both positive integers. $\square$

Strictly speaking the elements of an array can be anything at all. We shall restrict attention to arrays of numbers, which include as a special case functions of variables.

Example 1.1.1. Some 2×3 Matrices.
In general, 2×3 matrices have the form

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}.$$

Observe the use of subscripts to distinguish between different elements. In general, subscripts are arranged as $a_{row\ column}$. Also note that is standard practice to denote matrices by capital letters and individual elements of a matrix by a corresponding lower case letter. This is exactly the same as was done with sets except that now we use two subscripts to identify row and column position, whereas previously we only need a single subscript to distinguish the individual elements of a set. For example, we might arrange the set of numbers $\{3, 2, 8, 2, 5, 8\}$ in the array

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 8 \\ 2 & 5 & 8 \end{bmatrix}.$$

Similarly, we could define a matrix of functions of x of the form

$$\mathbf{F}(x) = \begin{bmatrix} x^2 & \ln x & g(x) \\ \sqrt{x} & x^2 + \ln x + \sqrt{x} & 8 \end{bmatrix}.$$ $\square$

There are alternative representations of matrices which define the matrix in terms of its typical element. From the previous example, the matrix $\mathbf{A}$ has typical element a_{ij} where i can take the values 1 or 2 and j can take any of the values 1, 2 or 3. Two other notations that you may encounter for $m \times n$ matrices are:

$$\mathbf{A} = \{a_{ij}\}_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}} \qquad \text{and} \qquad \mathbf{A} = [a_{ij}]_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}.$$

In situations where the dimensions are obvious from the context it is not uncommon for the subscripts to be omitted entirely; so, for example, we might write $\mathbf{A} = \{a_{ij}\}$. Such practice is, of course, making strong demands of the reader to keep track of dimensions which can lead to all sorts of sloppy mistakes; we shall avoid such definitions.

1.2 Special Cases

Within the set of all $m \times n$ matrices there are certain special cases of particular interest. The three criteria that have been used in assembling this list are (i) the structure of these matrices makes them particularly easy to work with, (ii) these matrices demonstrate the sense in which matrices are generalizations of ordinary numbers, or (iii) these matrices are encountered so frequently in either economics or statistics that one should be familiar with them. Although the following list is far from exhaustive it is a useful beginning.

Scalar Matrix A scalar matrix, or simply a scalar, is a matrix consisting of a single row and a single column. That is, a scalar is just a number, e.g. $m = n = 1: \ \mathbf{A} = a$. We shall use lower case letters to denote scalars.[1]

Row Vectors A matrix consisting of a single row is referred to as a row vector. That is, for $m = 1, n > 1$, $\mathbf{a} = [a_1 \ \ldots \ a_n]$ is a row vector. Note that vectors are typically denoted by a lower case letter, in this case $\mathbf{a}$. Two features help in distinguishing between vectors and scalars which are also typically denoted by lower case letters. First, elements of a matrix typically have subscripts whereas vectors usually don't, although there are some circumstances where they do. Second, like other matrices, we shall use a bold-faced type for vectors; that is, use $\mathbf{a}$ rather than a to denote a vector.

Column Vectors A matrix consisting of a single column is referred to as a column vector. That is, for $m > 1, n = 1$,

$$\mathbf{A} = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix}$$

is a column vector.

The standard practice when working with vectors is to define one type of vector as the default type. In this unit, and indeed most commonly in practice, you should assume that all vectors are column vectors unless they are explicitly defined to be row vectors. Subsequent unqualified use of the word 'vector', which will be the norm, should be read as 'column vector'.

[1] A scalar-valued function of a matrix (or set of numbers) is one that yields a scalar as its answer. For example, if a and y are scalars when $a = f(\mathbf{B})$ and $y = g(x_1, \ldots, x_n)$, then $f(\cdot)$ is a scalar-valued function of $\mathbf{B}$ and $g(\cdot)$ is a scalar-valued function of $x_1, \ldots, x_n$. Of course, functions $f(\cdot)$ of the form $y = f(x)$, where both x and y are scalars are trivial examples of scalar-valued functions. The alternative to a scalar-valued function of a matrix is a matrix-valued function of a matrix. For example, if $\mathbf{G} = h(\mathbf{X})$, where both $\mathbf{G}$ and $\mathbf{X}$ are matrices, then $h(\cdot)$ is a matrix-valued function.

Zero Matrix A matrix consisting entirely of zeros is referred to as a zero matrix, i.e. $a_{ij} = 0$ for all $i = 1, \ldots, m; j = 1, \ldots, n$. A zero matrix shall be denoted $\mathbf{0}_{m\times n}$ where the dimensions may be omitted if obvious from the context. In the context of matrix algebra it serves much the same role as does zero in regular number theory, viz. adding zero doesn't change anything whereas any product involving zero is zero.

Square Matrix As you might expect, a square matrix is one with the same number of rows as columns, i.e. $m = n$. Square matrices occur very frequently in economics and statistics and so are very important. Within the class of square matrices there is a number of special cases that are of particular interest.

Diagonal Matrix A diagonal matrix has the property that $a_{ij} = 0$ for all $i \neq j$.

Example 1.2.1. Diagonal Matrix.

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \qquad \square$$

All non-zero numbers of diagonal matrices appear on the *principal diagonal*. The principal diagonal is also sometimes referred to as the *leading diagonal*. Note that some, but not all, of the elements on the principal diagonal can take the value zero. Strictly, a zero matrix could be thought of as a diagonal matrix where all the values on the leading diagonal are also zero although this is not done in practice.

Identity Matrix One particularly important diagonal matrix is the identity matrix, typically denoted I_m or simply I if the dimension m is obvious from the context, in which all the elements of the leading diagonal are unity. Given that all elements of an identity matrix are known, the matrix is completely defined once its dimension or *order*, is known. That is, an identity matrix of dimension $m \times m$ is referred to as an identity matrix of order m.

Example 1.2.2. Identity Matrix of Order 3.

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \square$$

Symmetric Matrix Symmetric matrices are characterized by the property $a_{ij} = a_{ji}$ for all $i = 1, \ldots, m; j = 1, \ldots, n$. That is, one side of the principal (or leading) diagonal is the mirror image of the other.

Example 1.2.3. Symmetric Matrix.

$$\mathbf{S} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix} \qquad \square$$

Observe that all diagonal matrices are also symmetric. One important feature of symmetric matrices is that they have only $n(n+1)/2$ distinct elements, whereas an arbitrary $m \times n$ matrix has mn distinct elements. That is, in general, every element of an arbitrary matrix might be different from every other element in the matrix but for symmetric matrices this is not true by definition.

Triangular Matrix A triangular matrix is one for which either all the elements above the principal diagonal are zero, called a *lower triangular matrix*, or all the elements below the principal diagonal are zero, called an *upper triangular matrix*. Diagonal matrices are special cases which are both lower triangular and upper triangular.

Example 1.2.4. Triangular Matrices.

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix} \qquad \square$$

Lower Triangular Upper Triangular

In some circumstances upper and lower triangular matrices can be thought of as providing square roots of square matrices. This is a more advanced application than we require in this unit. However, we will see in Section 2 that matrix reduction is concerned with the creation of upper triangular matrices.

Vandermonde Matrix The Vandermonde matrix of order n takes the form

$$V_n = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-2} & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-2} & x_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{n-2} & x_n^{n-1} \end{bmatrix} \tag{1.2.1}$$

There are some variations on the theme that are also called Vandermonde matrices, e.g., $\widetilde{V}_n = \mathrm{diag}(x_1, \dots, x_n)V_n$, the essential feature is that the elements of each row belong to a geometric progression. Also, some definitions allow for the matrix to be rectangular although we will restrict attention to the case where V_n is square. The primary use of the Vandermonde matrix is in polynomial interpolation, although it also occurs in other circumstances.

1.3 Matrix Operations

Having defined matrices, an obvious question is what can be done with them. In this section we gather the fundamental matrix operations, this treatment is far from complete although sufficient for our purposes.

1.3.1 Matrix Transpose

Definition 1.3.1. Matrix Transpose.
If

$$\mathbf{A} = [a_{ij}]_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}$$

then the *transpose* of $\mathbf{A}$, denoted either $\mathbf{A}^\top$ or $\mathbf{A}'$, is

$$\mathbf{A}^\top = [a_{ji}]_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}. \qquad \square$$

Observe that if $\mathbf{A}$ is $m \times n$ then $\mathbf{A}^\top$ is $n \times m$.

Example 1.3.1. Matrix Transpose.

$$\text{If } \mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \text{ then } \mathbf{A}^\top = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}. \qquad \square$$

Note that $\left(\mathbf{A}^\top\right)^\top = \mathbf{A}$. Also, if $\mathbf{S}$ is a symmetric matrix then $\mathbf{S}^\top = \mathbf{S}$. Finally, the transpose of a lower triangular matrix is an upper triangular matrix and vice versa.

1.3.2 Scalar Multiplication

Definition 1.3.2. Scalar Multiplication.
Scalar multiplication is the operation of multiplying a matrix by a scalar (or number). If $\mathbf{A}$ is an $m \times n$ matrix and c is any scalar then the operation of scalar multiplication is defined as

$$c\mathbf{A} = [ca_{ij}]_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}. \qquad \square$$

Note that the order of scalar multiplication does not matter, so that $c\mathbf{A} = \mathbf{A}c$.

Example 1.3.2. Scalar Multiplication.

$$7\begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix} = \begin{bmatrix} 14 & 21 \\ 35 & 42 \end{bmatrix}. \qquad \square$$

A variety of different results can be obtained from scalar multiplication. In particular, for scalar values of c, c_1 and c_2,

(i) $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$

(ii) $c_1\mathbf{A} + c_2\mathbf{A} = (c_1 + c_2)\mathbf{A}$

(iii) $(c\mathbf{A})^\top = c\mathbf{A}^\top$

(iv) $0 \times \mathbf{A} = \mathbf{0}$

(v) $-1 \times \mathbf{A} = -\mathbf{A}$

Given that scalars admit interpretations as matrices this is a *very* special (and exceptional) case of matrix multiplication.

1.3.3 Matrix Addition

In order to add matrices they must be of the same dimension. That is, if one matrix has dimensions $m \times n$ then it can only be added to other matrices of dimensions $m \times n$.

Definition 1.3.3. Conformability Under Matrix Addition.
Matrices that are of the same dimension are said to be *conformable under (matrix) addition*.

Definition 1.3.4. Matrix Addition.
If $\mathbf{A}$ and $\mathbf{B}$ are both $m \times n$ matrices, so that they are conformable under matrix addition, then the operation of *matrix addition* is defined as $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where $c_{ij} = a_{ij} + b_{ij}$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$. $\square$

Definition 1.3.5. Matrix Subtraction.
The difference between two matrices $\mathbf{C} = \mathbf{A} - \mathbf{B}$ is defined to be $\mathbf{C} = \mathbf{A} + (-1 \times \mathbf{B})$, so that $c_{ij} = a_{ij} - b_{ij}$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$. $\square$

The difference between two matrices is actually the outcome of two operations, scalar multiplication and matrix addition, and so it also requires that the two matrices involved be conformable under matrix addition.

Example 1.3.3. Matrix Addition.
Let

$$\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} 11 & 12 & 13 \\ 24 & 25 & 26 \end{bmatrix},$$

then

(i) $3\mathbf{A} = \begin{bmatrix} 3 \times 1 & 3 \times 4 \\ 3 \times 2 & 3 \times 5 \\ 3 \times 3 & 3 \times 6 \end{bmatrix} = \begin{bmatrix} 3 & 12 \\ 6 & 15 \\ 9 & 18 \end{bmatrix}$;

(ii) $\mathbf{B} + \mathbf{C} = \begin{bmatrix} 1+11 & 2+12 & 3+13 \\ 4+24 & 5+25 & 6+26 \end{bmatrix} = \begin{bmatrix} 12 & 14 & 16 \\ 28 & 30 & 32 \end{bmatrix}$;

(iii) $\mathbf{C} - \mathbf{B} = \mathbf{C} + (-1) \times \mathbf{B} = \begin{bmatrix} 10 & 10 & 10 \\ 20 & 20 & 20 \end{bmatrix} = 10 \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix}$;

(iv) $\mathbf{A} + \mathbf{B}$ is not defined as $\mathbf{A}$ and $\mathbf{B}$ are not conformable under matrix addition. $\square$

1.3.4 Matrix Equality

Definition 1.3.6. Matrix Equality.
The statement $\mathbf{A} = \mathbf{B}$ requires that the matrices $\mathbf{A}$ and $\mathbf{B}$ be conformable under matrix addition, both $m \times n$ say, and that $a_{ij} = b_{ij}$ for all $i = 1, \ldots, m$ and $j = 1, \ldots n$. $\square$

1.3.5 Matrix Multiplication

Definition 1.3.7. Matrix Multiplication.
Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be a $p \times q$ matrix. The *matrix product* $\mathbf{C} = \mathbf{AB}$ is well defined if and only if $\mathbf{A}$ and $\mathbf{B}$ are conformable under matrix multiplication,[2] in which case $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$. $\mathbf{C}$ has dimension $m \times q$. □

Example 1.3.4. Matrix Multiplication.
Define

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}, \qquad \text{and} \qquad \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}.$$

If $\mathbf{C} = \mathbf{AB}$ then[3]

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31}, \\ c_{12} &= a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32}, \\ c_{21} &= a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31}, \\ c_{22} &= a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32}. \end{aligned}$$ □

In the definition of matrix multiplication (Definition 1.3.7) it was required that the two matrices be conformable under matrix multiplication. In the product $\mathbf{AB}$ this required that $\mathbf{A}$ have the same number of columns as $\mathbf{B}$ has rows. If, for example, $\mathbf{A}$ was 3×2 and $\mathbf{B}$ was 2×5 then the product $\mathbf{AB}$ is well-defined but the product $\mathbf{BA}$ is not; unlike scalar multiplication, the order of matrix multiplication matters. Consequently, a formal definition of conformability under matrix multiplication must take account of the order of multiplication. In the product $\mathbf{AB}$ we say that $\mathbf{A}$ *pre-multiplies* $\mathbf{B}$, and that $\mathbf{B}$ *post-multiplies* $\mathbf{A}$.

Definition 1.3.8. Conformability Under Matrix Multiplication.
Two matrices are *conformable under matrix multiplication* if the pre-multiplying matrix has the same number of columns as the post-multiplying matrix has rows. □

As an aside, even when both products $\mathbf{AB}$ and $\mathbf{BA}$ are well-defined, which implies that both $\mathbf{AB}$ and $\mathbf{BA}$ are square, cases where $\mathbf{AB} = \mathbf{BA}$ are the exception rather than the rule.

Example 1.3.5. Commutative Matrices.
If

$$\mathbf{A} = \begin{bmatrix} 8 & 6 \\ 4 & 2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 6 & 6 \\ 4 & 0 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 3 & 7 \\ 1 & 6 \end{bmatrix},$$

then $\mathbf{AB} = \mathbf{BA}$ but, for example, $\mathbf{AC} \neq \mathbf{CA}$. If $\mathbf{AB} = \mathbf{BA}$ then $\mathbf{A}$ and $\mathbf{B}$ are said to be *commutative under matrix multiplication*. For square matrices, the identity matrix and the zero matrix will always form part of a commutative pair, although they clearly need not be part of every commutative pair! □

[2] Definition 1.3.8 defines conformability under matrix multiplication. In this case, conformability requires that $(n = p)$.

[3] More succinct statements of these results involve sigma notation. Thus, $c_{11} = \sum_{k=1}^{3} a_{1k} b_{k1}$, $c_{12} = \sum_{k=1}^{3} a_{1k} b_{k2}$, $c_{21} = \sum_{k=1}^{3} a_{2k} b_{k1}$ and $c_{22} = \sum_{k=1}^{3} a_{2k} b_{k2}$. The notational economy of sigma notation in this context increases with k.

The notion of conformability under matrix multiplication reveals just how extraordinary is scalar multiplication of a matrix. Interpreting scalars as 1×1 matrices, the multiplication of a matrix by a scalar can be thought of as forming the product of two matrices that are not conformable under matrix multiplication, unless the matrix is either a scalar or a vector. Such an inconsistency means that matrix multiplication is not a perfect matrix analogue of scalar multiplication. There are, in fact, other sorts of matrix multiplication that seem more natural analogues of scalar multiplication, although they are beyond the scope of this course.

Rules of Matrix Multiplication

Below are some commonly used rules for, and special cases of, matrix multiplication. Unless otherwise stated assume appropriate conformabilities.

(i) Order matters! From Example 1.3.3, $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{BC}$ and $\mathbf{CA}$ are well-defined but $\mathbf{AC}$ and $\mathbf{CB}$ are not.

(ii) In general, $\mathbf{AB} \neq \mathbf{BA}$.

(iii) If $\mathbf{D} = \mathbf{AB}$ and $\mathbf{E} = \mathbf{BC}$, then $\mathbf{ABC} = \mathbf{DC} = \mathbf{AE}$.

(iv) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ and $(\mathbf{B} + \mathbf{C})\mathbf{A} = \mathbf{BA} + \mathbf{CA}$.

(v) If, for scalar c, $\mathbf{D} = c\mathbf{A}$ and $\mathbf{E} = c\mathbf{B}$, then $c(\mathbf{AB}) = \mathbf{DB} = \mathbf{AE}$.

(vi) $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$.

(vii) If $\mathbf{A}$ is $m \times n$ then $\mathbf{I}_m\mathbf{A} = \mathbf{AI}_n = \mathbf{A}$, $\mathbf{0}_{p\times m}\mathbf{A} = \mathbf{0}_{p\times n}$ and $\mathbf{A0}_{n\times q} = \mathbf{0}_{m\times q}$.

(viii) If $\mathbf{A}$ is a square matrix of order n then $\mathbf{A}^p = \underbrace{\mathbf{A} \cdot \mathbf{A} \cdots \mathbf{A}}_{p \text{ terms}}$. By definition, $\mathbf{A}^0 = \mathbf{I}_n$.

(ix) The matrices $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{AA}^\top$ are always symmetric.

Example 1.3.6. Interpreting Matrices: Factory Production.
Suppose that a factory produces three types of products, P_1, P_2 and P_3 say, and that four inputs are used in the production process, namely M_1, M_2, M_3 and M_4. Next suppose that each unit of P_1 requires 2 units of M_1, 3 units of M_2, 1 unit of M_3 and 12 units of M_4 to produce. Similarly, each unit of P_2 requires 7 units of M_1, 9 units of M_2, 5 units of M_3 and 20 units of M_4 to produce and each unit of P_3 requires 8 units of M_1, 12 units of M_2, 6 units of M_3 and 15 units of M_4 to produce. This information is presented in the matrix $\mathbf{Q}$ which gives inputs per unit of output. Observe that each column of $\mathbf{Q}$ corresponds to a given input and each row to a given output.

$$\begin{array}{cc} & \begin{array}{cccc} M_1 & M_2 & M_3 & M_4 \end{array} \\ \mathbf{Q} = & \left[\begin{array}{cccc} 2 & 3 & 1 & 12 \\ 7 & 9 & 5 & 20 \\ 8 & 12 & 6 & 15 \end{array}\right] \begin{array}{c} P_1 \\ P_2 \\ P_3 \end{array} \end{array}$$

M_1, M_2 M_3 and M_4 cost \$10, \$12, \$15 and \$20 per unit, respectively. We shall represent this information in the vector $\mathbf{C}$.

$$\mathbf{C} = \begin{bmatrix} 10 \\ 12 \\ 15 \\ 20 \end{bmatrix}$$

Suppose, finally, that a customer orders seven units of P_1, twelve units of P_2 and five units of P_3. If the customer's order is represented by the vector

$$\mathbf{P} = \begin{bmatrix} 7 \\ 12 \\ 5 \end{bmatrix},$$

evaluate and interpret (i) $\mathbf{P}^\top\mathbf{Q}$, (ii) $\mathbf{QC}$, and (iii) $\mathbf{P}^\top\mathbf{QC}$.

SOLUTION

(i) $\mathbf{P}^\top\mathbf{Q} = \begin{bmatrix} 7 & 12 & 5 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 & 12 \\ 7 & 9 & 5 & 20 \\ 8 & 12 & 6 & 15 \end{bmatrix} = \begin{bmatrix} 138 & 189 & 97 & 399 \end{bmatrix}.$
The vector $\mathbf{P}^\top\mathbf{Q}$ contains the quantities of M_1, M_2 M_3 and M_4 required to produce the amounts of P_1, P_2 and P_3 ordered by the customer. In particular, the factory will need 138 units of M_1, 189 units of M_2, etc.

(ii) $\mathbf{QC} = \begin{bmatrix} 2 & 3 & 1 & 12 \\ 7 & 9 & 5 & 20 \\ 8 & 12 & 6 & 15 \end{bmatrix} \begin{bmatrix} 10 \\ 12 \\ 15 \\ 20 \end{bmatrix} = \begin{bmatrix} 311 \\ 653 \\ 614 \end{bmatrix}.$
$\mathbf{QC}$ gives the vector of costs per unit of output for each of P_1, P_2 and P_3. Hence, it costs \$311 to produce each unit of P_1, \$653 to produce each unit of P_2 and \$614 to produce each unit of P_3.

(iii) Using the previous two results it would be equally valid to calculate $\mathbf{P}^\top\mathbf{QC}$ as either

$$\mathbf{P}^\top\mathbf{QC} = \begin{bmatrix} 138 & 189 & 97 & 399 \end{bmatrix} \begin{bmatrix} 10 \\ 12 \\ 15 \\ 20 \end{bmatrix} = 13,083.$$

or

$$\mathbf{P}^\top\mathbf{QC} = \begin{bmatrix} 7 & 12 & 5 \end{bmatrix} \begin{bmatrix} 311 \\ 653 \\ 614 \end{bmatrix} = 13,083.$$

Either way, $\mathbf{P}^\top\mathbf{QC}$ gives the total cost of producing the customer's order, which is evidently \$13,083. □

1.3.6 Trace of a (Square) Matrix

Definition 1.3.9. Trace of a (Square) Matrix.
If $\mathbf{A}$ is a square matrix then the trace of $\mathbf{A}$, denoted trace($\mathbf{A}$) or simply tr(A), is defined to be the sum of the elements on the leading diagonal of $\mathbf{A}$. In particular, if $\mathbf{A}$ is of dimension $m \times m$, with typical element a_{ij}, $i, j = 1, 2, \ldots, m$, then

$$\text{trace}(\mathbf{A}) = \textstyle\sum_{j=1}^{m} a_{jj}. \qquad \square$$

We typically omit the term 'square' when talking about the trace of a matrix because the operation is only defined o the set of square matrices.

Example 1.3.7. Trace of a Matrix.

$$\text{trace} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = 1 + 5 + 9 = 16. \qquad \square$$

1.3.7 Matrix Inverse

Definition 1.3.10. Matrix Inverse.
If $\mathbf{A}$ is a square matrix and there exists a (square) matrix $\mathbf{C}$ such that $\mathbf{CA} = \mathbf{I}$, then $\mathbf{C}$ is called the *inverse* of $\mathbf{A}$. The inverse of a matrix $\mathbf{A}$ is typically denoted $\mathbf{A}^{-1}$. $\square$

Definition 1.3.11. Non-singular Matrix.
Any matrix $\mathbf{A}$ that is invertible (has an inverse) is said to be *non-singular*. If $\mathbf{A}$ cannot be inverted it is said to be a *singular* matrix. $\square$

All non-singular matrices are square. As non-square matrices cannot be inverted in the sense of Definition 1.3.10, the adjective 'singular' is typically reserved for non-invertible square matrices.

It is often useful to be able to talk about individual elements of a matrix inverse. The usual notational convention is that if $\mathbf{A} = \{a_{ij}\}$ then $\mathbf{A}^{-1} = \{a^{ij}\}$, $i, j = 1, \ldots, n$ say. This is most easily illustrated in the following example of a 2×2 matrix.

Example 1.3.8. Notational Convention for Elements of $\mathbf{A}^{-1}$: The 2×2 Case.

$$\text{If} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \text{then} \quad \mathbf{A}^{-1} = \begin{bmatrix} a^{11} & a^{12} \\ a^{21} & a^{22} \end{bmatrix}.$$

The elements of $\mathbf{A}$ and $\mathbf{A}^{-1}$ are related by the matrix equality $\mathbf{AA}^{-1} = \mathbf{I}_2$, or equivalently by the set of equations

$$\begin{aligned} a_{11}a^{11} + a_{12}a^{21} &= 1, & a_{11}a^{12} + a_{12}a^{22} &= 0, \\ a_{21}a^{11} + a_{22}a^{21} &= 0, & a_{21}a^{12} + a_{22}a^{22} &= 1, \end{aligned} \tag{1.3.1}$$

and so it is reasonable to ask whether the elements of $\mathbf{A}^{-1}$ can be written in terms of $\mathbf{A}$ or not. The answer is yes, however, the relationships are quite complicated and become increasingly so as the dimensions of $\mathbf{A}$ increase. Equation (1.3.1) can be solved to yield[4]

$$\begin{aligned} a^{11} &= \frac{a_{22}}{a_{11}a_{22} - a_{12}a_{21}}, & a^{12} &= -\frac{a_{12}}{a_{11}a_{22} - a_{12}a_{21}}, \\ a^{21} &= -\frac{a_{21}}{a_{11}a_{22} - a_{12}a_{21}}, & a^{22} &= \frac{a_{11}}{a_{11}a_{22} - a_{12}a_{21}}, \end{aligned} \tag{1.3.2}$$

or, in matrix notation,

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \tag{1.3.3}$$

□

Observe that all of the values in Equation (1.3.2) depend upon the quantity $a_{11}a_{22} - a_{12}a_{21}$. This quantity is called the determinant of $\mathbf{A}$ and shall be encountered again in Section 3.1. For matrices of larger dimension it is not useful to write down these relationships due to their complexity.

Properties of Matrix Inverse

(i) $\mathbf{A}^{-1}$ is both the left-inverse and right-inverse of $\mathbf{A}$. That is, $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ and $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

(ii) $\mathbf{A}^{-1}$ is the unique inverse of $\mathbf{A}$; there is no other choice for $\mathbf{C}$ that satisfies $\mathbf{C}\mathbf{A} = \mathbf{I}$.

(iii) $\mathbf{A}$ is the inverse of $\mathbf{A}^{-1}$. That is, every non-singular matrix is the inverse of some other matrix.

(iv) $(\mathbf{A}^{-1})^\top = \left(\mathbf{A}^\top\right)^{-1}$.

(v) $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, provided $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ both exist. For example, $\mathbf{A}\mathbf{B}$ may be square even though neither $\mathbf{A}$ nor $\mathbf{B}$ is square; an important special case is when $\mathbf{A}$ is not square but $\mathbf{B} = \mathbf{A}^\top$.

(vi) If the non-singular matrix $\mathbf{A}$ is symmetric then so is $\mathbf{A}^{-1}$.

(vii) If the non-singular matrix $\mathbf{A}$ is upper (lower) triangular then so is $\mathbf{A}^{-1}$.

We have seen that matrix algebra is very similar to regular algebra in many respects; one can add, subtract and multiply matrices, there are analogues of zero and one, and inverses can be formed which are similar in spirit to reciprocals of numbers.[5] There are clearly differences in that matrix addition and multiplication require conformable matrices. Further, the only real number that cannot be inverted is zero whereas only non-singular matrices can be inverted. In the same way that matrix multiplication is substantially more complicated than scalar multiplication

[4] You should check that the values of Equation (1.3.2) actually satisfy Equation (1.3.1).

[5] For Arthur Cayley (1821–1895), the inventor of matrices, the fact that the commutative law of multiplication was not generally applicable proved a major conceptual hurdle to the development of matrix algebra.

so too is the matrix analogue of division. Scalar division can be thought of as scalar multiplication by a scalar inverse or reciprocal, *viz.*

$$\frac{a}{b} = a \times \frac{1}{b} = \frac{1}{b} \times a, \text{ valid for all } b \neq 0.$$

Imagine how such an expression might be interpreted if $a \equiv \mathbf{A}$ and $b \equiv \mathbf{B}$ are both matrices. The first issue that arises is how to interpret $1/b$. If it is to be interpreted as $\mathbf{B}^{-1}$ then any notion of division has immediately been restricted to b being a non-singular matrix. Such a restriction excludes all non-square matrices, and a large number of square matrices as well. Assuming that $1/b$ should be interpreted as $\mathbf{B}^{-1}$, it is not clear whether

$$\frac{a}{b} = \mathbf{AB}^{-1} \qquad \text{or} \qquad \frac{a}{b} = \mathbf{B}^{-1}\mathbf{A}.$$

Unless $\mathbf{A}$ and $\mathbf{B}^{-1}$ are commutative under matrix multiplication it is clear that these two different interpretations of a/b will yield different answers. In order for $\mathbf{A}$ and $\mathbf{B}^{-1}$ to be commutative it is necessary that they are both square. Of course, there is no reason for us to assume that $\mathbf{A}$ is square. Indeed this is probably unjustifiable as conformability under matrix multiplication merely requires that $\mathbf{A}$ have the same number of columns (or rows) as $\mathbf{B}$ to be able to form $\mathbf{AB}^{-1}$ (or $\mathbf{B}^{-1}\mathbf{A}$). Clearly, imposing a conformability condition restricts the set of feasible matrices $\mathbf{A}$, whereas in scalar division any real number could appear in the numerator. Nevertheless, either $\mathbf{AB}^{-1}$ or $\mathbf{B}^{-1}\mathbf{A}$, but not both unless $\mathbf{A}$ is square, might reasonably be thought of as being a matrix analogue of scalar division as it includes scalar division as a special case.

The discussion of the previous paragraph seeks to highlight how difficult is any notion of matrix division. Because of all the restrictions there is little point in thinking of matrix division as a well-defined operation. That is, one can pre-multiply or post-multiply by a matrix inverse but not divide by a matrix. In particular, *never* write either $\mathbf{B}/\mathbf{A}$ or $\frac{\mathbf{B}}{\mathbf{A}}$ if you mean either $\mathbf{A}^{-1}\mathbf{B}$ or $\mathbf{BA}^{-1}$. This is one of those instances where treating matrices in exactly the same way as you treat numbers leads to confusion and, in all probability, the wrong answer.

1.4 Matrices With Special Properties Under Various Matrix Operations

In what follows we will encounter a variety of different types of matrices that have special properties in certain circumstances. It is extremely helpful to be aware of the situations where these properties can be exploited.

1.4.1 Idempotents

The set of real numbers is defined by a number of properties, many of which are shared by the set of real matrices, division being a notable exception that is explored in Section 3.5. One of the properties is that of idempotency. An idempotent is something that, when applied under certain operations leaves the object of interest unchanged. In particular, among the set of real numbers there are two idempotents.

The first is the number zero, which is an idempotent under addition.[6] That is, you can add zero to any number, x say, and be left with x. The second idempotent is one, which is the idempotent under multiplication. Any number is returned unchanged when multiplied by unity.[7]

Matrix algebra doesn't have the operation of division, as such, but addition and multiplication are both well-defined matrix operations and there is an idempotent for each. Under matrix addition, the idempotent is a zero matrix of the same dimension as the object of interest. Thus, if $\mathbf{X}$ is an $m \times n$ matrix then the $m \times n$ zero matrix is the relevant idempotent, i.e., $\mathbf{X} + \mathbf{0} = \mathbf{X}$.

When matrix multiplication is the operation of interest then the relevant idempotent is an identity matrix of appropriate dimension. If the object of interest is rectangular rather than square, then the dimension of the idempotent will depend upon whether pre- or post-multiplication is being considered. That is, if $\mathbf{X}$ is an $m \times n$ matrix then

$$\mathbf{I}_m\mathbf{X} = \mathbf{X}\mathbf{I}_n = \mathbf{X}.$$

While analogies with the set of real numbers constitute a useful starting point when trying to understand matrices, it is the case that properties don't necessarily port across exactly. Indeed, among the set of square matrices is the class of idempotent matrices defined by the property $\mathbf{A}^2 = \mathbf{AA} = \mathbf{A}$. That is, any matrix $\mathbf{A}$ which returns itself when squared is said to be an *idempotent matrix*. It follows that if $\mathbf{A}$ is idempotent then, for any integer power n, $\mathbf{A}^n = \mathbf{A}$. The identity matrix is a special case of an idempotent matrix.

Idempotent matrices have lots of special properties, many of which which we are yet to define. For the record:

(i) All idempotent matrices are square.

(ii) All idempotent matrices are singular (non-invertible), except for identity matrices, which are their own inverse and hence non-singular.

(iii) Because all idempotent matrices are singular (non-invertible) except for identity matrices, the identity matrix is the only idempotent matrix with full rank.

(iv) The product of two idempotent matrices is also idempotent if the matrices commute in multiplication., i.e., if $\mathbf{AB} = \mathbf{BA}$. To see this observe that $(\mathbf{AB})(\mathbf{AB}) = \mathbf{A}(\mathbf{BA})\mathbf{B} = \mathbf{AABB} = \mathbf{AB}$, where the second last equality follows from commutativity under matrix multiplication of $\mathbf{A}$ and $\mathbf{B}$, and the final equality follows from the idempotency of each of $\mathbf{A}$ and $\mathbf{B}$.

(v) When an idempotent matrix is subtracted from an identity matrix the result is another idempotent matrix.

(vi) Symmetric idempotent matrices are also projection matrices. If $\mathbf{P}$ denotes a projection matrix then $\hat{\mathbf{y}} = \mathbf{Py}$ is a projection of $\mathbf{y}$ onto the space spanned by the columns of $\mathbf{P}$. That is, $\hat{\mathbf{y}}$ is formed as a linear combination, or weighted sum, of the columns of $\mathbf{P}$, where the weights are given by the elements of $\mathbf{y}$.

[6] Here subtraction can be thought of as adding the negative of a number and so doesn't need separate treatment.

[7] Equally, division can be thought of as multiplication by the reciprocal of a number, which is well-defined for all real numbers except zero.

(vii) The trace of an idempotent matrix is equal to its rank (which will be a natural number, e.g., 1,2,3,...).

(viii) The determinant of an idempotent matrix is zero, unless the matrix is an identity matrix, in which case its determinant is equal to unity.

(ix) The eigenvalues of idempotent matrices are all equal to either zero or one.

We will re-visit some of these properties further down the track.

Example 1.4.1. An Idempotent Projection Matrix.
Let $\mathbf{X}$ denote an $m \times n$ matrix with full column rank. Then the matrix $P_X = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is idempotent and so too is the matrix $M_X = \mathbf{I}_m - P_X$. (Prove it!) □

1.4.2 Orthogonal and Orthonormal Matrices

Orthogonality is a property of pairs of vectors. Suppose that we have two vectors, $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x}'\mathbf{x} = \xi \neq 0$ and $\mathbf{y}'\mathbf{y} = \psi \neq 0$. If $\mathbf{x}'\mathbf{y} = 0$ then we say that $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* vectors. If, in addition to being orthogonal, $\xi = \psi = 1$ then we say that the vectors are *orthonormal*. We can gather sets of mutually orthogonal/orthonormal vectors of the same order into matrices. We shall refer to such matrices as sub-orthogonal/orthonormal matrices unless they are square. The case of square orthogonal/orthonormal matrices is of particular interest but we shall illustrate one property that applies more generally first. Suppose that an $m \times n$ matrix $\mathbf{G}$ has rows that constitute a set of orthogonal vectors. In this case, $\mathbf{G}\mathbf{G}' = \mathrm{diag}(d_1, \ldots, d_m)$. However, it need not be the case that $\mathbf{G}'\mathbf{G}$ is diagonal. For example, let

$$\mathbf{G} = \begin{bmatrix} 2 & 0 & 6 \\ 0 & 3 & 0 \end{bmatrix}.$$

Then

$$\mathbf{G}\mathbf{G}' = \begin{bmatrix} 40 & 0 \\ 0 & 9 \end{bmatrix} \quad \text{but} \quad \mathbf{G}'\mathbf{G} = \begin{bmatrix} 4 & 0 & 12 \\ 0 & 9 & 0 \\ 12 & 0 & 36 \end{bmatrix}.$$

We see that $\mathbf{G}\mathbf{G}'$ is diagonal but $\mathbf{G}'\mathbf{G}$ is not (in this case, although it may be in others, depending on the values of the elements of $\mathbf{G}$.). In the same way, if $\mathbf{G}$ was comprised of orthogonal columns then we know that $\mathbf{G}'\mathbf{G}$ will be diagonal, but that may not be the case for $\mathbf{G}\mathbf{G}'$. Similarly, if the orthogonal vectors in $\mathbf{G}$ all have unit length then our diagonal matrices reduce to identity matrices of appropriate order. For example, if

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \implies \mathbf{G}\mathbf{G}' = \mathbf{I}_2 \quad \text{but} \quad \mathbf{G}'\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \neq \mathbf{I}_3.$$

As another example, that doesn't look quite so special, let

$$\mathbf{U} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \implies \mathbf{U}\mathbf{U}' = \mathbf{I}_2 \quad \text{but} \quad \mathbf{U}'\mathbf{U} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} & \frac{1}{3} \\ -\frac{1}{6} & \frac{5}{6} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \neq \mathbf{I}_3.$$

However, when an orthogonal/orthonormal matrix, $\mathbf{G}$ say, is square and of order m, then it satisfies

$$\mathbf{G}\mathbf{G}' = \mathbf{G}'\mathbf{G} = \mathbf{I}_m.$$

In particular, this result implies that $\mathbf{G}^{-1} = \mathbf{G}'$. By way of example,

$$\mathbf{G} = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \sqrt{2} \\ \sqrt{3} & -\sqrt{3} & 0 \\ 1 & 1 & -2 \end{bmatrix}$$

provides an example of an orthogonal matrix, where $\mathbf{G}\mathbf{G}' = \mathbf{G}'\mathbf{G} = \text{diag}(6,6,6)$. The matrix $\mathbf{H} = \frac{1}{\sqrt{6}}\mathbf{G}$ is an example of an orthonormal matrix, i.e. $\mathbf{H}\mathbf{H}' = \mathbf{H}'\mathbf{H} = \mathbf{I}_3$. A second example is readily available if one augments the matrix $\mathbf{U}$ from above to obtain

$$\widetilde{\mathbf{U}} = \begin{bmatrix} & \mathbf{U} & \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{bmatrix} \Longrightarrow \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}' = \widetilde{\mathbf{U}}'\widetilde{\mathbf{U}} = \mathbf{I}_3.$$

As a pair of remarks on nomenclature, first, people often make no distinction between orthogonal and orthonomal matrices and will use the terminology 'orthogonal matrix' to mean an orthonormal matrix as described above. In a nod to common usage, we may find ourselves doing the same as we move forward. Second, when we speak of a sub-orthogonal matrix we will mean a subset of columns (or rows) from an orthogonal matrix. For example, $\mathbf{U}$ is a matrix that is both sub-orthogonal (because the row vectors are orthogonal to one another) and sub-orthonormal (because the individual row vectors have unit length). It is worth being aware that given a sub-orthogonal matrix, $\mathbf{H}_1$ say, it is always possible to find a matrix $\mathbf{H}_2$ such that $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2]$ is orthogonal. (We have stated this in terms of columns of a matrix but an equivalent statement can be made for rows, as we did in the construction of $\widetilde{\mathbf{U}}$ given $\mathbf{U}$.) Indeed, there is an infinitude of such matrices $\mathbf{H}_2$, if orthogonality is all that is required, although the dimension of the set of possibilities is reduced if orthonormality is required as well.

Finally, a very important property of orthogonal/orthonormal matrices is that products of them are also orthogonal/orthonormal. For example, if $\mathbf{A}$ and $\mathbf{B}$ are both orthonormal matrices of order m, say, so that $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}_m$ and $\mathbf{B}'\mathbf{B} = \mathbf{B}\mathbf{B}' = \mathbf{I}_m$, then

$$(\mathbf{AB})'(\mathbf{AB}) = \mathbf{B}'\mathbf{A}'\mathbf{A}\mathbf{B} = \mathbf{B}'\mathbf{I}_m\mathbf{B} = \mathbf{B}'\mathbf{B} = \mathbf{I}_m$$

and

$$(\mathbf{AB})(\mathbf{AB})' = \mathbf{A}\mathbf{B}\mathbf{B}'\mathbf{A}' = \mathbf{A}\mathbf{I}_m\mathbf{A}' = \mathbf{A}\mathbf{A}' = \mathbf{I}_m,$$

as required.

Permutation Matrices

Among the class of orthonormal matrices sits that of *permutation matrices*. A permutation matrix of order m is an identity matrix with two of its rows interchanged. Because it is so closely related to an identity matrix it follows that permutation matrices are also square. We shall denote such matrices by the symbol $\mathbf{E}$ because we will encounter them when we deal with elementary operations. Moreover, we shall use subscripts to denote which rows have been swapped. For example, $\mathbf{E}_{12}$ has its first and second row interchanged. If we let $m = 3$ then we see that

$$\mathbf{E}_{12} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Observe that $\mathbf{E}_{12}$ is symmetric about its leading diagonal. This is true of permutation matrices in general. Another important property of permutation matrices is that they are orthonormal, which means that $\mathbf{E}_{rs}\mathbf{E}'_{rs} = \mathbf{E}'_{rs}\mathbf{E}_{rs} = \mathbf{I}_m$. Note that, by symmetry, $\mathbf{E}'_{rs} = \mathbf{E}_{rs}$ and so both the equalities of the previous sentence reduce to $\mathbf{E}_{rs}\mathbf{E}_{rs} = \mathbf{I}_m$. If you stop and think about this for an instant then it immediately becomes obvious why this is so. Return to our example of $\mathbf{E}_{12}$ above. We obtained it by permuting the first two rows of $\mathbf{I}_3$, $\mathbf{E}_{12}\mathbf{I}_3 = \mathbf{E}_{12}$ if you like. If you premultiply $\mathbf{E}_{12}$ by itself then we are permuting these two rows again, i.e., putting them back where they started, which brings us back to the original identity matrix.

A very important property of these matrices, true of orthogonal matrices more broadly, is that they are non-singular and, hence invertible. Indeed, due to the symmetry of permutation matrices, they are their own inverse, as we saw in the previous paragraph.

As one might expect, the role of permutation matrices is to re-order the elements of other matrices. Specifically, suppose that we have a pemutation matrix $\mathbf{E}$ then premultiplication by $\mathbf{E}$ will permute the rows of a matrix whereas post-multiplication by $\mathbf{E}$ will permute the columns of a matrix. For example, using our earlier definition of $\mathbf{E}_{12}$,

$$\mathbf{E}_{12} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 4 & 5 & 6 \\ 1 & 2 & 3 \\ 7 & 8 & 9 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 4 & 5 & 6 \\ 1 & 2 & 3 \\ 7 & 8 & 9 \end{bmatrix} \mathbf{E}_{12} = \begin{bmatrix} 5 & 4 & 6 \\ 2 & 1 & 3 \\ 8 & 7 & 9 \end{bmatrix}.$$

Note that the subscripts on $\mathbf{E}$ remain the same, with the impact of permuting rows or columns determined by whether you pre- or post-multiply by the permutation matrix.

1.5 Linearly Independent Vectors

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ denote a set of p non-null $n \times 1$ vectors. We say that the p vectors are *linearly independent* if there exists no set of weights $\omega_1, \ldots, \omega_p$ such that

$$\omega_1\mathbf{x}_1 + \cdots + \omega_p\mathbf{x}_p = \mathbf{0},$$

except for $\omega_1 = \omega_2 = \cdots \omega_p = 0$. Note that there can be as many zeros as you wish among the weights provided that not all of them are zero. Vectors that are not linearly independent are said to be *linearly dependent*. That is, if there exists a set of weights $\omega_1, \ldots, \omega_p$ that are not all zero such that

$$\omega_1\mathbf{x}_1 + \cdots + \omega_p\mathbf{x}_p = \mathbf{0},$$

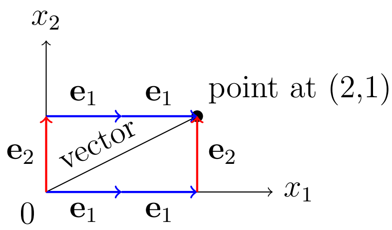


Figure 1.1: An Orthogonal Basis

then the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ are linearly dependent.

Null vectors are never included as part of a set of linearly independent vectors. This is to preclude situations like the following. Suppose that, in our set of vectors, $\mathbf{x}_p = \mathbf{0}$. Further suppose that all of the weights are zero except for ω_p, so that

$$0\mathbf{x}_1 + \cdots + 0\mathbf{x}_{p-1} + \omega_p\mathbf{0} = \mathbf{0}, \qquad \omega_p \neq 0.$$

This might lead one to conclude that the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{p-1}$ are linearly dependent, when in fact this equation is really silent on that question, because

$$\begin{aligned} & 0\mathbf{x}_1 + \cdots + 0\mathbf{x}_{p-1} + \omega_p\mathbf{0} = \mathbf{0} \\ \Longrightarrow\ & 0\mathbf{x}_1 + \cdots + 0\mathbf{x}_{p-1} = (1-\omega_p)\mathbf{0} = \mathbf{0} \end{aligned}$$

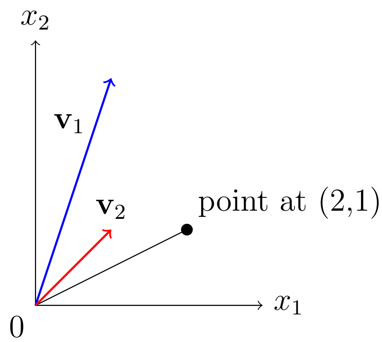
if it were the case that these were the only weights satisfying this last equation then that would imply linear independence of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{p-1}$, but we don't (and can't) know that given this information. For this reason the null vector $\mathbf{0}$ is always deemed to belong to a set of linearly dependent vectors.

Sets of linearly independent vectors are very important in many aspects of matrix (or linear) algebra. Every vector of m real numbers, $\mathbf{v}_j$, say can be thought of as both a point in n-dimensional space, denoted $\mathbb{R}^m$, and also a vector joining the origin to that point. For example, we can think of a vector $\mathbf{x} = [2,1]'$ being both the point (2,1) in the x_1-x_2 plane and also the vector joining the origin to that point, see Figure 1.1. If we think of the vectors $\mathbf{e}_1 = [1,0]'$ and $\mathbf{e}_2 = [0,1]'$ (represented in Figure 1.1 by blue and red arrows, respectively) then we can reach the point (2,1) by either adding the vectors $\mathbf{e}_1 + \mathbf{e}_1 + \mathbf{e}_2$, which takes us along the x_1 axis and then up parallel to the x_2 axis, or up the x_2 axis by $\mathbf{e}_2$ and then to the right parallel to the x_1 axis by twice the $\mathbf{e}_1$ vector.[8] The vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ form a *basis set* for this two dimensional space. We say that $\mathbf{e}_1$ and $\mathbf{e}_2$ span the space $\mathbb{R}^2$. That is, every point (ω_1, ω_2) in the x_1-x_2 plane can be reached as $\omega_1\mathbf{e}_1 + \omega_2\mathbf{e}_2$ and it is this feature that makes $\mathbf{e}_1$ and $\mathbf{e}_2$ a basis set.

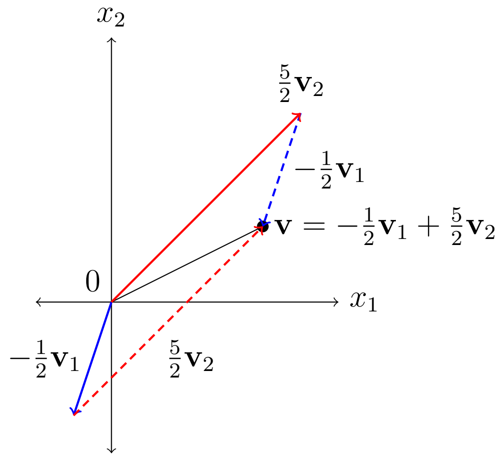
Now, basis sets are not unique. Any pair of vectors $\mathbf{v}_1 = [v_{11}, v_{21}]'$ and $\mathbf{v}_2 = [v_{12}, v_{22}]'$ that intersect at a single point (so they can't be parallel, nor can they be co-linear) can be used as a basis set of $\mathbb{R}^2$. For example, consider the vectors $\mathbf{v}_1 = [1,3]'$ and $\mathbf{v}_2 = [1,1]'$, which are represented in Figure 1.2a by blue and red arrows, respectively. These two form a basis for the x_1-x_2 plane. For instance, the vector $\mathbf{v} = [2,1]'$ can be obtained as the following combination of $\mathbf{v}_1$ and $\mathbf{v}_2$ (see Figure 1.2b):

$$\mathbf{v} = -\tfrac{1}{2}\mathbf{v}_1 + \tfrac{5}{2}\mathbf{v}_2.$$

[8] If you were feeling contrary then you may also travel the path $\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_1$, but that would clutter the graph too much to bother drawing.



(a) The Basis Vectors



(b) Spanning the Space

Figure 1.2: A Non-Orthogonal Basis

We see that it doesn't matter whether you arrive at $\mathbf{v}$ by adding an image of $\mathbf{v}_2$ to the end of $\mathbf{v}_1$ or by adding an image of $\mathbf{v}_1$ to the end of $\mathbf{v}_2$. The images of vectors are represented by dashed lines of the appropriate colour. The relationship of all this to matrix algebra is that we can stack our vectors in a matrix: $[\mathbf{e}_1, \mathbf{e}_2] = \mathbf{I}_2$ so that our basis set is comprised of the columns of an identity matrix. These vectors form an orthonormal basis set. Equally, we might write $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2]$. These vectors may or may not be orthogonal and they may or may not have unit length. No basis set is either right or wrong, it is just that some can be more convenient to work with than others. An orthnormal basis set constructed from the columns of an identity matrix is often an easy one to work with as it simply corresponds to the Cartesian coordinate system that you have been using all your life. But the set of axes represented by $\mathbf{v}_1$ and $\mathbf{v}_2$ in Figure 1.2a works just as effectively at spanning the x_1-x_2 plane as does x_1-x_2 axes.

More generally, in order to span $\mathbb{R}^m$ it is necessary to have a set of m linearly independent m-vectors. The columns of $\mathbf{I}_m$ stand out as obvious candidates. If you have p m-vectors, with $p > m$, then only m of them can be linearly independent and the remaining $p - m$ can be written as linear combinations of the members of the basis set. Note that in such situations, there may be many different ways of constructing a suitable basis set. Similarly, if $p < m$ then, at most, your p vectors can only be spanning a p-dimensional sub-space of $\mathbb{R}^m$ (and less than p if there are linear dependencies among your vectors). For example, if the only vector you have is $\mathbf{e}_1$ then you are only able to span those points which lie along the x_1 axis, and so on.

There is one very important aspect of linearly independent vectors and matrices of which you should be aware:

Theorem 1.1 (Linear Independence and Matrices)**.** *The number of linearly independent rows in a matrix is the same as the number of linearly independent columns.* □

This will become clear when we consider elementary row and column operations in the next chapter; see also the discussion of rank in Section 2.5 and that of equivalent canonical form of Section 7.1.

We have seen that if we have a set of linearly independent vectors that we can stack in a matrix $\mathbf{A} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$, where $\mathbf{A}$ is $m \times n$, $m \geq n$, then it means that there exists no vector $\mathbf{c}$ of dimension $n \times 1$ such that $\mathbf{Ac} = \mathbf{0}$ other than $\mathbf{c} = \mathbf{0}$. Moreover, the columns of $\mathbf{A}$ form a basis for an n-dimensional subset of $\mathbb{R}^m$, $m \geq n$. But this basis does not allow us to say anything about the rest of $\mathbb{R}^m$, which is a space of dimension $m - n$. As we shall see, when we are concerned with finding solutions to homogeneous equations of the form $\mathbf{Ax} = \mathbf{0}$, it is exactly this subset of $\mathbb{R}^m$ that contains our solutions and so it is of interest to us. Because it is of interest to us, this subset of $\mathbb{R}^m$ has a special name. It is known as the *null space* of $\mathbf{A}$ and denoted $\mathcal{N}(\mathbf{A})$, or somtimes $\text{Null}(\mathbf{A})$. Some books also call this the *kernel* of $\mathbf{A}$.

Not every matrix is comprised of linearly independent columns and rows, although every matrix other than a zero matrix will have at least some linearly independent columns and rows. If we call this number r for now, then it can be shown that the dimension of the null space of an arbitrary $m \times n$ matrix $\mathbf{A}$ is given by $m - r$. That is, a basis set for $\mathcal{N}(\mathbf{A})$ will require $m - r$ linearly independent vectors of dimension $m \times 1$. Although not stated as such, this result is, in essence, what is known as the Fundamental Theorem of Linear Algebra. It will be encountered again in Section 5.4, when we are interested in determining the number of linearly independent solutions to a system of linear equations.

1.6 An Introduction to Solving Matrix Equations

One of the great strengths of matrices is the economy of notation they afford when dealing with systems of linear equations which, in turn, are extremely important in both economics and statistics. Indeed, the principal application of the matrix inverse is that of solving linear equations, equations of the form $\mathbf{\Gamma X} = \mathbf{B}$ where $\mathbf{\Gamma}$ and $\mathbf{B}$ are matrices of constants and $\mathbf{X}$ is a matrix of variables.[9] In particular, if

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \ldots & \ldots & \ldots \\ \gamma_{n1} & \cdots & \gamma_{nn} \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix},$$

then $\mathbf{\Gamma X} = \mathbf{B}$ is a neat way of expression the system of n equations in the n unknowns $x_1, \ldots, x_n$

$$\begin{gathered} \gamma_{11}x_1 + \cdots + \gamma_{1n}x_n = b_1 \\ \vdots \\ \gamma_{n1}x_1 + \cdots + \gamma_{nn}x_n = b_n. \end{gathered}$$

Note that this system contains, as a special case, a system of two equations in two unknowns such as

$$\begin{aligned} 6x + 7y &= 20 \\ x - y &= -1. \end{aligned}$$

[9]In what follows, we shall always treat $\mathbf{B}$ and $\mathbf{X}$ as being vectors even though we will tend to retain the upper case notation that indicates that they may be matrices with more than one column. The reason for this is simply that the results we will obtain are equally valid whether $\mathbf{B}$ and $\mathbf{X}$ are vectors or not.

Here

$$\mathbf{\Gamma} = \begin{bmatrix} 6 & 7 \\ 1 & -1 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 20 \\ -1 \end{bmatrix}.$$

Rather than complicating the analysis with application specific detail, in what follows we shall consider the general problem of solving systems of linear equations.

Not all systems of linear equations admit a solution. However, it is possible to characterize those situations where the system has a unique solution.

Definition 1.6.1. Unique Solution to a System of Linear Equations.
If $\mathbf{\Gamma X} = \mathbf{B}$ then, provided that $\mathbf{\Gamma}$ is non-singular, $\mathbf{X} = \mathbf{\Gamma}^{-1}\mathbf{B}$ is the unique solution for $\mathbf{X}$, where the uniqueness of the solution for $\mathbf{X}$ follows from the uniqueness of a matrix inverse. □

Establishing this result is trivial on pre-multiplying both sides of the equation $\mathbf{\Gamma X} = \mathbf{B}$ by $\mathbf{\Gamma}^{-1}$ and recognizing that $\mathbf{\Gamma}^{-1}\mathbf{\Gamma} = \mathbf{I}$.

Example 1.6.1. Unique Solution to a System of Linear Equations.
If

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 22 \end{bmatrix},$$

solve for x_1 and x_2.

SOLUTION

Assuming the inverse exists the solution is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix}^{-1} \begin{bmatrix} 11 \\ 22 \end{bmatrix}.$$

The problem is finding the inverse. Using Equation (1.3.3) we have

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix}^{-1} = \frac{1}{2 \times 7 - 1 \times 3} \begin{bmatrix} 7 & -1 \\ -3 & 2 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 7 & -1 \\ -3 & 2 \end{bmatrix}.$$

We need to check that this inverse is correct. Thus,

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \times \frac{1}{11} \begin{bmatrix} 7 & -1 \\ -3 & 2 \end{bmatrix} = \mathbf{I}_2,$$

as required. Substituting for the inverse yields

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 7 & -1 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} 11 \\ 22 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$

It is always a good idea to check solutions, hence

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 11 \\ 22 \end{bmatrix},$$

as required. □

One thing this example illustrates is that having first to obtain a matrix inverse before being able to solve for $\mathbf{X}$ is an extremely cumbersome procedure which, as it transpires, is also unnecessary. Consider how we might approach this problem in the absence of matrices. The system of equations that we have to solve is

$$2x_1 + x_2 = 11 \tag{1.6.1}$$
$$3x_1 + 7x_2 = 22. \tag{1.6.2}$$

Multiplying Equation (1.6.1) by three and Equation (1.6.2) by two yields

$$6x_1 + 3x_2 = 33 \tag{1.6.3}$$
$$6x_1 + 14x_2 = 44. \tag{1.6.4}$$

Subtracting Equation (1.6.3) from Equation (1.6.4) yields

$$11x_2 = 11 \tag{1.6.5}$$

or, on dividing both sides of Equation (1.6.5) by eleven,

$$x_2 = 1. \tag{1.6.6}$$

On subtracting $14x_2$ from both sides of Equation (1.6.4), we see that

$$6x_1 = 44 - 14x_2 = 30, \tag{1.6.7}$$

where the final equality follows from Equation (1.6.6). Dividing both sides of Equation (1.6.7) by six yields the final solution of

$$x_1 = 5. \tag{1.6.8}$$

The values $x_1 = 5$ and $x_2 = 1$ are exactly the same as those obtained in Example 1.6.1. Obviously it is not necessary to calculate a matrix inverse to solve linear equations. The way we have solved these equations is to go through a sequence of transformations of the original system of equations, as stated in Equations (1.6.1) and (1.6.2), until we reduced that system to an equivalent system of equations, given by Equations (1.6.6) and (1.6.8), where we could read off the solutions directly. Our task now is to translate this operation into matrix algebra, a process called matrix reduction by elementary row operations.

To see what is involved consider the system of equations in matrix notation.

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 22 \end{bmatrix}.$$

Now, the first step adopted above was to multiply the first equation by three while leaving the second equation unchanged. This is achieved by pre-multiplying both sides of the matrix equation by a diagonal matrix as follows

$$\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 11 \\ 22 \end{bmatrix},$$

which yields

$$\begin{bmatrix} 6 & 3 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 33 \\ 22 \end{bmatrix}.$$

Notice the effect of pre-multiplication by a diagonal matrix. Every element of the first row of the matrix of coefficients is multiplied by three, which is the first element on the leading diagonal of the diagonal matrix and, similarly, elements of the second row of the coefficient matrix are multiplied by the value of the second element on the leading diagonal of the diagonal matrix. (What do you think would be the effect of post-multiplying by a diagonal matrix?) In the same way, we can multiply the second equation by two. Thus,

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 6 & 3 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 33 \\ 22 \end{bmatrix},$$

which yields

$$\begin{bmatrix} 6 & 3 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 33 \\ 44 \end{bmatrix}.$$

Observe that these two transformations could have been achieved in a single step because

$$\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.$$

Our next step is to subtract the first equation from the second. Because there is an interaction between equations our transformation will involve pre-multiplication by a matrix with non-zero off-diagonal elements. Thus,

$$\begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 6 & 3 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 33 \\ 44 \end{bmatrix},$$

which yields

$$\begin{bmatrix} 0 & 11 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 44 \end{bmatrix}.$$

Our next step was to divide the elements of the top row by eleven which requires pre-multiplication by a diagonal matrix.

$$\begin{bmatrix} \frac{1}{11} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 11 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{11} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 11 \\ 44 \end{bmatrix},$$

or

$$\begin{bmatrix} 0 & 1 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 44 \end{bmatrix}.$$

Our penultimate step was to subtract fourteen times x_2 from the bottom row of this matrix equation. This is done as follows

$$\begin{bmatrix} 1 & 0 \\ -14 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -14 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 44 \end{bmatrix},$$

giving

$$\begin{bmatrix} 0 & 1 \\ 6 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 30 \end{bmatrix}.$$

Finally, we need to divide the bottom row by six

$$\begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{6} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 6 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{6} \end{bmatrix} \begin{bmatrix} 1 \\ 30 \end{bmatrix},$$

giving

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix},$$

or

$$\begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}.$$

An interesting thing has happened. Pre-multiplying the vector $[x_1, x_2]^\top$ by the matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

has resulted in the order of the elements being reversed but the elements themselves were otherwise unchanged. Hence, we could do this again

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 5 \end{bmatrix},$$

to obtain

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$

There is a number of points to be observed from this rather long calculation. First, each of our transformations has taken us from one system of equations to a different but equivalent system of equations, where equivalent means that the systems of equations have the same solution. It may seem that this has been a more complicated task than just using the matrix inverse. Indeed, when working with 2×2 matrices, where we can write down the formula for the matrix inverse, this is probably true. However, it becomes increasingly less true as the dimensions of the problem (the number of equations and unknowns to be solved for) increase.

Second, if we write

$$\mathbf{T}_1 = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{T}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \ldots \mathbf{T}_k = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

then we could represent the series of transformations used above to obtain our solution as

$$\mathbf{T}_k \times \cdots \times \mathbf{T}_1 \boldsymbol{\Gamma} \mathbf{X} = \mathbf{T}_k \times \cdots \times \mathbf{T}_1 \mathbf{B}.$$

Note that, because our transformations involved pre-multiplication, the transformation matrices start next to $\boldsymbol{\Gamma}$ (and $\mathbf{B}$) and move out further to the left, so that $\mathbf{T}_k$ appears first on the line because it was the last transformation. Clearly, because our solution is of the form $\mathbf{X} = \mathbf{C}$, where $\mathbf{C}$ is some vector,

$$\mathbf{T}_k \times \cdots \times \mathbf{T}_1 \boldsymbol{\Gamma} = \mathbf{I},$$

which implies that

$$\mathbf{T}_k \times \cdots \times \mathbf{T}_1 = \mathbf{\Gamma}^{-1},$$

and also implies that

$$\mathbf{T}_k \times \cdots \times \mathbf{T}_1\mathbf{B} = \mathbf{\Gamma}^{-1}\mathbf{B}.$$

This is exactly what was given in Definition 1.6.1 as the unique solution to a system of equations. Thus, as an aside, we have developed a way of finding a matrix inverse, where it exists, via a series of matrix multiplications rather than by solving the equation $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

Finally, you may have noticed that the vector $\mathbf{X}$ has remained unchanged throughout all of this, the transformations effected $\mathbf{\Gamma}$ and $\mathbf{B}$. For this reason, people have developed a simplified notation that omits $\mathbf{X}$ entirely. Before introducing this notation we will look more closely at the sorts of transformations that can be used to move between equivalent systems of equations.

Chapter 2

Elementary Matrix Operations and Matrix Reduction

2.1 Reduction of Matrices

At the end of the previous section we illustrated how to solve a system of equations using elementary row operations. These operations were used to transform a system of equations into an equivalent system of equations where the arbitrary coefficient matrix of the original system had been reduced to an identity matrix in the final system. Let us now formalize some of these ideas.

Definition 2.1.1. A Reduced Matrix.
A matrix is said to be a *reduced matrix*, or in *reduced row echelon form*, if:

(i) The first non-zero entry of any row, called the leading entry, is unity and all other entries in the column of the leading entry are zeros.

(ii) The leading entry in each row is to the right of the leading entry of each row above it.

(iii) Any rows consisting entirely of zeros are at the bottom of the matrix. □

Example 2.1.1. Reduced Row Echelon Form of a Matrix.
Which of these four matrices are in reduced row echelon form?

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 7 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 3 & 0 & 2 \\ 0 & 1 & 7 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & 1 & 7 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

A and **D** are both in reduced row echelon form, the latter case illustrates that there can be columns of zeros before encountering the first non-zero element. **B** is not in reduced form because the leading element in the first row is not unity. **C** is not in reduced form because there is a row of zeros above a row with non-zero elements. □

2.2 Elementary Matrix Operations

Definition 2.2.1. Elementary Row Operations.
Matrices are reduced to row echelon form by *elementary row operations*, which

consist of (i) interchanging rows of a matrix, (ii) multiplying a row of a matrix by a non-zero scalar, and (iii) adding a multiple of one row of a matrix to a multiple of other rows of that matrix. $\square$

The following notation is adopted to indicate which row operation has been applied.[1]

Interchange Rows $R_i \leftrightarrow R_j$ indicates that the ith and jth rows have been interchanged.

Scale Rows $R_i^* = \alpha R_i$ indicates that the ith row has been multiplied by the scale factor α.

Weighted Sums of Rows $R_i^* = \alpha_1 R_1 + \ldots + \alpha_m R_m$ $(\alpha_i \neq 0)$ indicates that the ith row has been replaced by a weighted sum of all m rows of the matrix. The restriction $\alpha_i \neq 0$ has been imposed because otherwise it would be easy to 'lose' one of the rows. If this happens the reduced form matrix that you finally arrive at will not be the reduced form of the matrix you started with. That is, all original m rows must be retained in some form throughout the row operations — you can't lose rows!

Example 2.2.1. Matrix Reduction by Elementary Row Operations. Reduce to row echelon form the matrix

$$\begin{bmatrix} 2 & 1 & 11 \\ 3 & 7 & 22 \end{bmatrix}.$$

SOLUTION

$$\begin{aligned}
\begin{bmatrix} 2 & 1 & 11 \\ 3 & 7 & 22 \end{bmatrix} \quad & \xrightarrow{R_1^* = R_2 - R_1} \quad \begin{bmatrix} 1 & 6 & 11 \\ 3 & 7 & 22 \end{bmatrix} \\
& \xrightarrow[R_2^* = R_2 - 3R_1]{} \quad \begin{bmatrix} 1 & 6 & 11 \\ 0 & -11 & -11 \end{bmatrix} \\
& \xrightarrow[R_2^* = -\frac{1}{11} R_2]{R_1^* = R_1 + \frac{6}{11} R_2} \quad \begin{bmatrix} 1 & 0 & 5 \\ 0 & 1 & 1 \end{bmatrix}
\end{aligned}$$

The horizontal arrows indicate that we are transforming to an *equivalent* matrix. The sense in which it is equivalent shall be defined later. $\square$

Example 2.2.2. Reduction to Row Echelon Form (A Second Example). Problem: Transform the matrix $\mathbf{C}$ to reduced row echelon form, where

$$\mathbf{C} = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 1 & -1 \\ 3 & 2 & -1 \end{bmatrix}.$$

Solution:

$$\begin{bmatrix} 2 & -1 & 1 \\ 1 & 1 & -1 \\ 3 & 2 & -1 \end{bmatrix} \xrightarrow{R_1 \leftrightarrow R_2} \begin{bmatrix} 1 & 1 & -1 \\ 2 & -1 & 1 \\ 3 & 2 & -1 \end{bmatrix} \xrightarrow[R_3^* = R_3 - 3R_1]{R_2^* = R_2 - 2R_1} \begin{bmatrix} 1 & 1 & -1 \\ 0 & -3 & 3 \\ 0 & -1 & 2 \end{bmatrix}$$

[1] Note the use of the * superscript to indicate transformed rows. This is not really standard notations but it does make clear exactly what is going on.

$$\xrightarrow{R_2^*=-R_2/3} \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & 2 \end{bmatrix} \xrightarrow{R_3^*=R_3+R_2} \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}. \qquad \square$$

Comments on the previous example:

- With two more operations we could have reduced $\mathbf{C}$ to an identity matrix:

$$\begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow[R_2^*=R_2+R_3]{R_1^*=R_1-R_2} \mathbf{I}_3.$$

- In terms of the steps from the previous slide, converting the first non-zero entry in a row to unity is called a *pivot transformation* or, simply, *pivoting*.
- Converting to zeros the elements of lower rows in the pivot's column is called *sweeping the column*.

Although we have used annotated arrows to represent the various row operations, the operations themselves are best thought of as premultiplications by non-singular matrices, matrices that we will call *elementary matrices* because they capture elementary operations. Note that if we premultiplied by singular matrices then that is the same as losing information that cannot be recovered. Such transformations are invalid elementary operations. Let us re-visit the previous example and make this explicit.

Example 2.2.3. Example 2.2.2 Re-Visited.
In Example 2.2.2, our first elementary row operation we to interchange the first and second rows (in order to get a 1 in the top right corner). We represented that step by

$$\begin{bmatrix} 2 & -1 & 1 \\ 1 & 1 & -1 \\ 3 & 2 & -1 \end{bmatrix} \xrightarrow{R_1 \leftrightarrow R_2} \begin{bmatrix} 1 & 1 & -1 \\ 2 & -1 & 1 \\ 3 & 2 & -1 \end{bmatrix}.$$

Instead, we might have premultiplied by the matrix

$$\mathbf{E}_{12} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus,

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 1 \\ 1 & 1 & -1 \\ 3 & 2 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 \\ 2 & -1 & 1 \\ 3 & 2 & -1 \end{bmatrix}$$

So, the elementary operation of interchanging rows (columns) of some matrix $\mathbf{C}$, say, amounts to the pre- (post-) multiplication of $\mathbf{C}$ by a permutation matrix. The other elementary matrix operations involve the scaling of rows (columns) and replacing rows (columns) by linear combinations of rows (columns), linear combinations that include the initial row (column), so that no information is ever lost. These operations can also be capture by pre- (post-) multiplication by a properly constructed non-singular matrix. The previous example contains example of both. Specifically, the second operation performed, that of sweeping out the remaining elements of the first column so that we are left with zeros below the first row, was represented by

$$\begin{bmatrix} 1 & 1 & -1 \\ 2 & -1 & 1 \\ 3 & 2 & -1 \end{bmatrix} \xrightarrow[R_3^*=R_3-3R_1]{R_2^*=R_2-2R_1} \begin{bmatrix} 1 & 1 & -1 \\ 0 & -3 & 3 \\ 0 & -1 & 2 \end{bmatrix}.$$

Instead, we could represent this pair of operations by the matrix

$$\mathbf{E}_2 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}$$

i.e.,

$$\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 \\ 2 & -1 & 1 \\ 3 & 2 & -1 \end{bmatrix} \xrightarrow[R_3^*=R_3-3R_1]{R_2^*=R_2-2R_1} \begin{bmatrix} 1 & 1 & -1 \\ 0 & -3 & 3 \\ 0 & -1 & 2 \end{bmatrix}.$$

We see that the second row of $\mathbf{E}_2$ corresponds to the row operation $R_2^* = R_2 - 2R_1$ and the third row to $R_3^* = R_3 - 3R_1$. Interestingly, the first row of $\mathbf{E}_2$ is just the first row of an identity matrix. This row is here to keep $\mathbf{E}_2$ square and non-singular but in such a way that the first row of the target matrix is left untouched. Throughout the solution there are other examples of replacing rows by linear combinations of rows in order to sweep out columns in rows below the leading value of unity. For example, the relevant matrices at the fourth and fifth steps are

$$\mathbf{E}_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \qquad \mathbf{E}_5 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

respectively, where $\mathbf{E}_4$ took us to reduced row echelon form and then $\mathbf{E}_5$ reduced the resulting matrix to an identity. It will be left as an exercise to convince yourself that this is true.

At the third step of the solution we scaled a row by $-\frac{1}{3}$, in a pivoting operation. The matrix for this operation, $\mathbf{E}_3$ is mostly comprised of an identity matrix, because we are only operating on a single row. The exception being that the second element on the leading diagonal is $-\frac{1}{3}$. This means that the operation applies to the second row and the second row alone. That is,

$$\begin{bmatrix} 1 & 1 & -1 \\ 0 & -3 & 3 \\ 0 & -1 & 2 \end{bmatrix} \xrightarrow{R_2^*=-R_2/3} \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

becomes

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 \\ 0 & -3 & 3 \\ 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & 2 \end{bmatrix}. \qquad \square$$

There is a couple of interesting features of this example. First, the number of steps taken to reach the final solution is more than required. This was so that each step clearly illustrated the row operation in question. So we note that we could have achieved reduced row echelon form in a single step by pre-multiplying $\mathbf{C}$ by a single matrix, $\mathbf{K}$ say, where

$$\mathbf{K} = \mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_{12} = \begin{bmatrix} 0 & 1 & 0 \\ -\frac{1}{3} & \frac{2}{3} & 0 \\ -\frac{1}{3} & -2\frac{1}{3} & 1 \end{bmatrix},$$

where the order in which the elementary operator matrices appear in the product reflects the fact that each subsequent row operation involves pre-multiplication. If we further pre-multiply $\mathbf{K}$ by $\mathbf{E}_5$ we obtain

$$\mathbf{E}_5\mathbf{K} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ -\frac{2}{3} & -1\frac{2}{3} & 1 \\ -\frac{1}{3} & -2\frac{1}{3} & 1 \end{bmatrix}.$$

But we saw that

$$\mathbf{E}_5\mathbf{K}\mathbf{C} = \mathbf{E}_5\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_{12}\mathbf{C} = \mathbf{I}_3$$

Now we know that any matrix $\mathbf{A}$ which premultiplies another matrix $\mathbf{C}$ to yield an identity matrix must be the inverse of $\mathbf{C}$, i.e., $\mathbf{A} = \mathbf{C}^{-1}$. That is,

$$\mathbf{C}^{-1} = \mathbf{E}_5\mathbf{K} = \mathbf{E}_5\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_{12} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 0 \\ -2 & -5 & 3 \\ -1 & -7 & 3 \end{bmatrix},$$

which demonstrates that we can assemble the inverse of a matrix as a product of elementary row operating matrices. Let us formalize this result.

Theorem 2.1 (Matrix Inversion Via Elementary Row Operations)**.** *If* $\mathbf{C}$ *is an* $n \times n$ *non-singular matrix then there exists a sequence of elementary row operations* $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_q$ *such that*

$$\mathbf{C}^{-1} = \mathbf{E}_q \times \cdots \times \mathbf{E}_1, \quad q \leq n. \qquad \square$$

So far, the examples given here have focussed on elementary row operations but we might equally have focussed on elementary column operations. The ideas are the same.

Definition 2.2.2. Elementary Column Operations.
Matrices are reduced to column echelon form by *elementary column operations*, which consist of (i) interchanging columns of a matrix, (ii) multiplying a column of a matrix by a non-zero scalar, and (iii) adding a multiple of one column of a matrix to a multiple of other columns of that matrix. $\square$

The following notation is adopted to indicate which column operation has been applied.[2]

Interchange Columns $C_i \leftrightarrow C_j$ indicates that the ith and jth columns have been interchanged.

[2]Note the use of the * superscript to indicate transformed columns. This is not really standard notations but it does make clear exactly what is going on.

Scale Column $C_i^* = \alpha C_i$ indicates that the ith column has been multiplied by the scale factor α.

Weighted Sums of Column $C_i^* = \alpha_1 C_1 + \ldots + \alpha_n C_n$ $(\alpha_i \neq 0)$ indicates that the ith column has been replaced by a weighted sum of all n columns of the matrix. The restriction $\alpha_i \neq 0$ has been imposed because otherwise it would be easy to 'lose' one of the columns. If this happens the reduced form matrix that you finally arrive at will not be the reduced form of the matrix you started with. That is, all original n columns must be retained in some form throughout the column operations — you can't lose columns!

Example 2.2.4. Finding an Inverse Via Elementary Column Operations. Here we will find $\mathbf{C}^{-1}$ via elementary column operations, where $\mathbf{C}$ is as defined in the previous two examples. One possible path is the following:

(i) Interchange the first and third columns of $\mathbf{C}$, to put a 1 in the top left corner, according to

$$\mathbf{C}\mathbf{E}_{13} = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 1 & -1 \\ 3 & 2 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 1 & 1 \\ -1 & 2 & 3 \end{bmatrix} = \mathbf{C}_1 \quad \text{(say)}.$$

(ii) Sweep out the first row of $\mathbf{C}_1$:

$$\mathbf{C}_1\mathbf{E}_2 = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 1 & 1 \\ -1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 3 \\ -1 & 1 & 5 \end{bmatrix} = \mathbf{C}_2 \quad \text{(say)}.$$

(iii) Interchange the second and third columns to get a non-zero element on the leading diagonal:

$$\mathbf{C}_2\mathbf{E}_3 = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 3 \\ -1 & 1 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 3 & 0 \\ -1 & 5 & 1 \end{bmatrix} = \mathbf{C}_3 \quad \text{(say)}.$$

(iv) Scale the $(2,2)$ element of $\mathbf{C}_3$ to unity and also sweep out the sub-diagonal elements.

$$\mathbf{C}_3\mathbf{E}_4 = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 3 \\ -1 & 1 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 \\ -\frac{2}{3} & -\frac{5}{3} & 1 \end{bmatrix} = \mathbf{I}_3.$$

We conclude that

$$\mathbf{C}^{-1} = \mathbf{E}_{13}\mathbf{E}_2\mathbf{E}_3\mathbf{E}_4 = \frac{1}{3}\begin{bmatrix} 1 & 1 & 0 \\ -2 & -5 & 3 \\ -1 & -7 & 3 \end{bmatrix},$$

as before. Note that the only reason that this process took one fewer steps than we had with the elementary row operations previously is that we combined two steps into one in the final step of this process.

As a final example, it is worth noting that in many situations one is going to want to combine both row and column operations. For example, in the first comment on Example 2.2.2 we noted that we could use row operations to sweep out the columns in the reduced row echelon form to reduce the matrix $\mathbf{C}$ to an identity matrix. In Example 2.2.3 this final step was represented by the matrix $\mathbf{E}_5$. However, this was only possible because the reduced row echelon form was of full *column* rank. This is not always the case, as is illustrated by Example 2.2.1, where we see that here is no combination of row operations that would allow us to sweep out the rows to the right of the left-most ones in each row. The reason for this is that the reduced row echelon form does not have full column rank. In this particular example the reduced row echelon form has full row rank but, in general, that need not be the case as the reduced row echelon form may contain rows of zeros. It is elementary column operations that allow to sweep out rows.

Example 2.2.5. Combining Elementary Row and Column Operations. Our starting point will be the matrix that was reduced in Example 2.2.1. That is, let

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 11 \\ 3 & 7 & 22 \end{bmatrix}.$$

Then the elementary row operations required to reduce $\mathbf{A}$ to row echelon form can be captured in the matrix

$$\mathbf{P} = \begin{bmatrix} 1 & \frac{6}{11} \\ 0 & -\frac{1}{11} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 7 & -1 \\ -3 & 2 \end{bmatrix}.$$

That is,

$$\mathbf{PA} = \frac{1}{11} \begin{bmatrix} 7 & -1 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 & 11 \\ 3 & 7 & 22 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 5 \\ 0 & 1 & 1 \end{bmatrix}.$$

There is no elementary row operations that will allow us to reduce the elements of the third column of this matrix to a column of zeros without introducing non-zero numbers to those elements that are already zero. Nevertheless, this is relatively easy to do using elementary column operations. In particular, we see that there exists a matrix $\mathbf{Q}$ such that

$$\mathbf{PAQ} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{2.2.1}$$

One possible choice for $\mathbf{Q}$ is

$$\mathbf{Q} = \begin{bmatrix} -4 & -5 & -5 \\ -1 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix}.$$

If you find it harder to work out the elementary column operations in your head than you do the elementary row operations then don't worry, you are not alone. But remember, the matrix algebra is just solving systems of linear equations, about which much more will be said shortly. The point being that you can always expand things out into a more familiar form in order to proceed. Writing

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}$$

we see that (2.2.1) corresponds to the set of equations

$$
\begin{aligned}
1 \times q_{11} + 0 \times q_{21} + 5 \times q_{31} &= 1 \\
1 \times q_{12} + 0 \times q_{22} + 5 \times q_{32} &= 0 \\
1 \times q_{13} + 0 \times q_{23} + 5 \times q_{33} &= 0 \\
0 \times q_{11} + 1 \times q_{21} + 1 \times q_{31} &= 0 \\
0 \times q_{12} + 1 \times q_{22} + 1 \times q_{32} &= 1 \\
0 \times q_{13} + 1 \times q_{23} + 1 \times q_{33} &= 0
\end{aligned}
$$

From these 6 equations we find the following 6 results:

$$
\begin{aligned}
q_{11} &= 1 - 5q_{31}, & q_{12} &= -5q_{32}, & q_{13} &= -5q_{33}, \\
q_{21} &= -q_{31}, & q_{22} &= 1 - q_{32}, & q_{23} &= -q_{33},
\end{aligned}
$$

so that

$$
\mathbf{Q} = \begin{bmatrix} 1 - 5q_{31} & -5q_{32} & -5q_{33} \\ -q_{31} & 1 - q_{32} & -q_{33} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}.
$$

The choice for $\mathbf{Q}$ given above corresponds to $q_{31} = q_{32} = q_{33} = 1$. But any other choice would do, except $q_{31} = q_{32} = q_{33} = 0$, as in this case $\mathbf{Q}$ is singular (because it has both a row and a column of zeros).[3] $\square$

2.3 Solving Systems by Reduction When the Solution is Unique

Consider the system of equations $\mathbf{\Gamma X} = \mathbf{B}$, where $\mathbf{\Gamma}$ is an $n \times n$ matrix and both $\mathbf{X}$ and $\mathbf{B}$ have n rows and an arbitrary number of columns. The matrix $\mathbf{\Gamma}$ is called the *coefficient matrix* of the system and the matrix $\mathbf{A} = [\mathbf{\Gamma}|\,\mathbf{B}]$ is called the *augmented coefficient matrix* of the system.[4] The augmented coefficient matrix completely describes the system of equations which can be 'solved' for $\mathbf{X}$ by reducing the augmented coefficient matrix to row echelon form. As an a

Example 2.3.1. Reducing the Augmented Coefficient Matrix.
Recall, from Example 1.6.1, the matrix equation

$$
\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 22 \end{bmatrix}.
$$

[3] Remember that the matrices capturing elementary column operations, like those capturing elementary row operations, must have full rank, so that $\mathbf{Q}$ non-singular. Also note that the dimensions of $\mathbf{Q}$ must match those of $\mathbf{A}$, not those of $\mathbf{P}$ and so, unless $\mathbf{A}$ is square, $\mathbf{P}$ and $\mathbf{Q}$ will have different dimensions.

[4] The augmented coefficient matrix has been defined with a vertical bar to separate the various components. Although commonly done for expository purposes it is not intrinsic to the definition and so we could equally write $\mathbf{A} = [\mathbf{\Gamma}\ \mathbf{B}]$. It is also common to see dashed or dotted lines used in place of the vertical bar, e.g. $\mathbf{A} = [\mathbf{\Gamma} \vdots \mathbf{B}]$. Because it helps to illustrate where the various components of the augmented coefficient matrix come from, we will tend to retain the vertical bar but it is not strictly required.

The coefficient and augmented coefficient matrices for this system are

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \qquad \text{and} \qquad \left[\begin{array}{cc|c} 2 & 1 & 11 \\ 3 & 7 & 22 \end{array}\right],$$

respectively. From the previous example the augmented coefficient matrix reduces to row echelon form as

$$\left[\begin{array}{cc|c} 2 & 1 & 11 \\ 3 & 7 & 22 \end{array}\right] \longrightarrow \left[\begin{array}{cc|c} 1 & 0 & 5 \\ 0 & 1 & 1 \end{array}\right]. \qquad \square$$

Recall that the horizontal arrow implied *equivalence*. They are equivalent matrices in the sense that they completely describe systems of equations which have the same solution. That is,

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 22 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$$

both have the same solution; namely, $x_1 = 5$ and $x_2 = 1$.

Recall, from Definition 1.6.1, that when $\mathbf{\Gamma}$ is non-singular the unique solution to the Equation $\mathbf{\Gamma X} = \mathbf{B}$ is $\mathbf{X} = \mathbf{\Gamma}^{-1}\mathbf{B}$. From this it follows that if the augmented coefficient matrix is $[\mathbf{\Gamma}|\,\mathbf{B}]$ and $\mathbf{\Gamma}$ is non-singular then the reduced row echelon form of this matrix must be $\left[\,\mathbf{I}\,|\,\mathbf{\Gamma}^{-1}\mathbf{B}\right]$. Setting $\mathbf{B} = \mathbf{I}$ yields a simple method for obtaining the inverse of an arbitrarily large matrix, because $\mathbf{\Gamma}^{-1}\mathbf{I} = \mathbf{\Gamma}^{-1}$; that is, if it exists, $\mathbf{\Gamma}^{-1}$ can be found by finding the reduced row echelon form of the augmented coefficient matrix $[\mathbf{\Gamma}|\,\mathbf{I}]$. This must work because, by the definition of a matrix inverse, $\mathbf{\Gamma X} = \mathbf{I}$ implies that $\mathbf{X} = \mathbf{\Gamma}^{-1}$.

Example 2.3.2. Inverting a Matrix by Row Reduction.
In an earlier example we found

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix}^{-1} = \frac{1}{11} \begin{bmatrix} 7 & -1 \\ -3 & 2 \end{bmatrix}.$$

The same results could have been obtained by first constructing the augmented matrix

$$\left[\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 3 & 7 & 0 & 1 \end{array}\right],$$

and then reducing this matrix to row echelon form as follows:

$$\left[\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 3 & 7 & 0 & 1 \end{array}\right] \xrightarrow[R_2^*=2R_2-3R_1]{R_1^*=R_2-R_1} \left[\begin{array}{cc|cc} 1 & 6 & -1 & 1 \\ 0 & 11 & -3 & 2 \end{array}\right] \xrightarrow[R_2^*=\frac{1}{11}R_2]{R_1^*=R_1-\frac{6}{11}R_2} \left[\begin{array}{cc|cc} 1 & 0 & \frac{7}{11} & -\frac{1}{11} \\ 0 & 1 & -\frac{3}{11} & \frac{2}{11} \end{array}\right].$$

$\square$

2.4 Solving Systems by Reduction: The General Case

2.4.1 Introduction

In the previous section we considered situations where there existed a unique solution to the system of equations $\mathbf{\Gamma X} = \mathbf{B}$. In Section 2.3 it was observed a linear system of equations was completely described by the augmented coefficient matrix and so these other cases must be able to be cast in terms of the augmented coefficient matrix. Further, the equivalence of systems of equations allows us to restrict attention to reduced form matrices without loss of generality.

The case of a unique solution correspondeds to the case where the coefficient matrix, $\mathbf{\Gamma}$ say, was invertible. The next step is to explore cases where $\mathbf{\Gamma}$ is not non-singular. This includes situations where $\mathbf{\Gamma}$ is not square and where $\mathbf{\Gamma}$ is square but singular. In what follows we shall assume that $\mathbf{\Gamma}$ is $m \times n$, $\mathbf{B}$ is $m \times 1$ and $\mathbf{A}$ is $m \times (n + 1)$. Note that non-singular $\mathbf{\Gamma}$ requires $m = n$.

It is common practice to decompose equations of the form $\mathbf{\Gamma X} = \mathbf{B}$ into two classes; namely, homogeneous and non-homogeneous equations.

Definition 2.4.1. Homogeneous Equations.
The linear system of equations $\mathbf{\Gamma X} = \mathbf{B}$ is said to be *homogeneous* if $\mathbf{B} = \mathbf{0}$ and *non-homogeneous* otherwise. □

If a system of linear equations is homogeneous there is always at least one solution, namely $\mathbf{X} = \mathbf{0}$. This is known as the *trivial* solution. Beyond the trivial solution the distinction is not relevant. In the examples that follow certain values of $\mathbf{B}$ are referred to generically as b_1 and b_2, these values are free to take any value including zero. Thus, the discussion that follows incorporates both homogeneous and non-homogeneous systems of equations.

2.4.2 More Equations Than Variables

The first case that shall be considered is where $m > n$, which means that there are more equations than variables. To focus discussion we shall explore the particular case where $m = 3$ and $n = 2$. In this case, the reduced form augmented coefficient matrix has the general structure:

$$\mathbf{A} = \left[\begin{array}{cc|c} 1 & 0 & b_1 \\ 0 & 1 & b_2 \\ 0 & 0 & b_3 \end{array}\right]. \tag{2.4.1}$$

That is, in this system there are three equations and two variables. The first equation implies that $x_1 = b_1$ and the second equation implies $x_2 = b_2$. The third equation implies that $b_3 = 0$. But remember that the elements of $\mathbf{X}$ are the only variables in our equations, the elements of $\mathbf{\Gamma}$ and $\mathbf{B}$ are parameters (or constants). Consequently there are two cases to consider, the first is when $b_3 = 0$ is true, so that the third equation is *consistent* with the facts, and the other is when $b_3 \neq 0$, so that the third equation is *inconsistent* with the facts.

If $b_3 = 0$ then Equation (2.4.1) contains a row of zeros, however a unique solution for $\mathbf{X}$ can be found from the remaining two rows. This means that there was one *redundant* equation in the original system. A redundant equation is one that provides no additional information that can assist in solving for $\mathbf{X}$ given the other equations in the system. Another way of saying this is that *each* of the equations could be written as a weighted sum of the other two, so that only two of the original equations contain distinct information about $\mathbf{X}$. The choice of the redundant equation in the original system is completely arbitrary; any one of the equations could be ignored and the same solution for $\mathbf{X}$ obtained. One reason for working with the reduced augmented coefficient matrix is that redundant equations are reduced to rows of zeros and so no time need be wasted trying to solve those equations.

In general, if *any* $(m - n)$ of the original equations are weighted sums of the other n equations they will generate $(m - n)$ rows of zeros in $\mathbf{A}$. In such cases $(m - n)$ of the rows are redundant and a unique solution can be obtained based on an appropriately chosen set of n original equations.

In equation (2.4.1), the implied solution for the third equation is $b_3 = 0 \times x_1 + 0 \times x_2 = 0$. Any value for b_3 other than $b_3 = 0$ yields an inconsistency. That is, there are no choices for either x_1 or x_2 that satisfy this equation if $b_3 \neq 0$.

In general, if you have $(m - n)$ more equations than variables then, unless there are also $(m - n)$ redundant equations, there is no possible solution that satisfies all of the equations. In this case the equations are said to be *inconsistent*.

2.4.3 Fewer Equations Than Variables

A system of equations with fewer equations than variables corresponds to the case where $m < n$. Within this case there are two sub-cases. First, we shall consider the case where $(n - m)$ of the columns are zero and then we will allow the excess columns to contain non-zero elements.

Over-Parametrized Systems

The first sub-case to be considered is when $n > m$ but where there are $(n - m)$ columns of zeros. For example, suppose

$$\mathbf{A} = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 7 \\ 0 & 1 & 0 & 8 \end{array}\right]. \tag{2.4.2}$$

The column of zeros in the coefficient matrix is a reflection of the fact that the original system has been over-parametrized. In this case there are three variables but the system is inherently two dimensional in that it has nothing to say about the values that could be taken by the third variable. The only sensible course of action in this situation is to re-parametrize the model in terms of a smaller set of variables where a solution might be feasible.

Example 2.4.1. An Over-Parametrized System of Equations.
The matrix $\mathbf{A}$ in Equation (2.4.2) could have been obtained by trying to solve the equations

$$\begin{aligned} 4x_1 - 3x_2 &= 4 \\ x_1 + x_2 &= 15, \end{aligned}$$

for x_1, x_2 and x_3. The system should be thought of as determining x_1 and x_2 alone as it provides no information about the value of x_3. □

Correctly Parametrized Systems

As shown in the context of over-parametrized systems, columns of zeros in other than the last column of the augmented coefficient matrix are simply a reflection of trying to do something silly. Consequently, we will restrict attention to where all the columns of the augmented coefficient matrix contain non-zero elements. Again it is useful to cast the discussion in terms of a concrete example, hence let $m = 2$, $n = 3$ and

$$\mathbf{A} = \left[\begin{array}{ccc|c} 1 & 0 & 6 & 7 \\ 0 & 1 & 5 & 8 \end{array}\right].$$

In this system there are two equations and three variables. The implied solutions are

$$\begin{aligned} x_1 \quad + 6x_3 &= 7 \\ x_2 + 5x_3 &= 8 \end{aligned}$$

or

$$\begin{aligned} x_1 &= 7 - 6x_3 \\ x_2 &= 8 - 5x_3 \end{aligned} \tag{2.4.3}$$

Equation (2.4.3) is called a *parametric* solution because it depends upon the unknown parameter x_3. Although x_3 is typically thought of as a variable there is insufficient information in the system to determine its value. Furthermore, the values taken by the other variables depend upon the value that it takes. That is, once a value for x_3 is fixed (or made constant like a parameter) the system can be solved for the remaining variables. The solutions arising from any given value for x_3 are known as *particular* solutions. There is potentially an infinite number of particular solutions possible if, for example, x_3 can take any real value.

Example 2.4.2. A Particular Solution.
The particular solution to equation (2.4.3) corresponding to $x_3 = 4$ is $x_1 = -17$, $x_2 = -12$ and $x_3 = 4$. □

In general, the solution may involve as many parameters as there are insufficient equations to solve the system completely.

2.4.4 A Singular Coefficient Matrix

The final case to be considered is where the coefficient matrix is square ($m = n$) but singular. That is, the reduced form of $\mathbf{\Gamma}$ differs from $\mathbf{I}$. For example, suppose that $m = n = 3$ and let

$$\mathbf{A} = \left[\begin{array}{ccc|c} 1 & 0 & 2 & b_1 \\ 0 & 1 & 1 & b_2 \\ 0 & 0 & 0 & b_3 \end{array}\right].$$

In this system there are three equations and three variables but given any two of the equations the remaining equation contains no distinct information. That is, each of the equations can be written as a weighted sum of the other two. There are two possible cases. First, if $b_3 = 0$ then there is a parametric solution, and consequently an infinite number of particular solutions. Indeed the structure of this case is exactly the same as was seen in Section 2.4.3 where there were more variables than equations. Second, if $b_3 \neq 0$ then there is an inconsistency as the third equation implies that $b_3 = 0$. The structure of this case is identical to that of Section 2.4.2 where it was seen that inconsistent equations have no solution.

2.5 Rank of a Matrix

2.5.1 Definition and Notation

The discussion of Section 2.4 characterized all of the different situations that may be encountered but is extremely cumbersome. A more economical characterization is afforded by the notion of rank.

Definition 2.5.1. Rank of a Matrix.
The *rank* of a matrix $\mathbf{Z}$, denoted $\rho(\mathbf{Z})$ or $\rho_{\mathbf{Z}}$, is the number of non-zero rows in its reduced row echelon form.[5] □

Let $\mathbf{Z}$ denote an $m \times n$ matrix then rank has the following properties:

(i) $\rho(\mathbf{Z}) \leq \min(m, n)$.

(ii) If $\rho(\mathbf{Z}) = m$ then $\mathbf{Z}$ is said to have *full row rank*.

(iii) If $\rho(\mathbf{Z}) = n$ then $\mathbf{Z}$ is said to have *full column rank*.

(iv) If $\rho(\mathbf{Z}) = m = n$, so that $\mathbf{Z}$ is square, then $\mathbf{Z}$ is said to have *full rank*. Non-singular matrices have full rank, singular matrices don't.

Example 2.5.1. Determining the Rank of a Matrix.
Determine the ranks of the following matrices:

$$\mathbf{I}_3, \quad \mathbf{Z_1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{Z_2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{Z_3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{Z_4} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

SOLUTION

$$\rho\left(\mathbf{I}_3\right) = 3, \qquad \rho(\mathbf{Z_1}) = 2, \qquad \rho(\mathbf{Z_2}) = 2, \qquad \rho\left(\mathbf{Z_3}\right) = 1, \qquad \rho(\mathbf{Z_4}) = 2.$$

That is, $\mathbf{I}_3$ is non-singular, $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are singular but neither $\mathbf{Z_3}$ nor $\mathbf{Z_4}$ are square so that their singularity, or otherwise, is not an issue. □

We have defined rank above in terms of the reduced row echelon form. There are various other definitions available, all of which are equivalent in the sense that they all yield the same result. Here is one of them.

Definition 2.5.2. An Alternative Definition of Rank.
The rank of a matrix is equal to the number of linearly independent rows/columns that it contains. □

Recall from Theorem 1.1 that the number of linearly independent rows of a matrix is always the same as the number of linearly independent columns that it contains.

[5]Note that ρ is the Greek symbol rho and not a 'p' of any description.

2.5.2 Characterizing Solutions to Systems of Equations in Terms of Rank

Using the notion of rank it is possible to summarize the various cases that may be encountered in attempting to solve systems of linear equations. In what follows $\mathbf{\Gamma}$ and $\mathbf{A}$ shall denote the coefficient matrix and the augmented coefficient matrix of the system of equations, respectively.

(i) A unique solution for n variables requires n distinct consistent equations. That is, uniqueness requires $\rho(\mathbf{A}) = \rho(\mathbf{\Gamma}) = n$.

(ii) If you have more than n distinct equations then there is no solution. That is, if $\rho(\mathbf{A}) > \rho(\mathbf{\Gamma})$ there is no solution.

(iii) If you have fewer than n distinct equations there is potentially an infinite number of solutions. That is, if $\rho(\mathbf{A}) = \rho(\mathbf{\Gamma}) < n$ there is potentially an infinite number of solutions.

Note that $\rho(\mathbf{A}) < \rho(\mathbf{\Gamma})$ is impossible.

Chapter 3

Determinants and Cramer's Rule

A determinant is a scalar-valued function of a square matrix. Simply put, if $\mathbf{A}$ is an arbitrary square matrix then the determinant of $\mathbf{A}$, denoted $|\mathbf{A}|$ or $\det(\mathbf{A})$, is a number, regardless of the dimensions of $\mathbf{A}$. All of the other matrix operations that have been encountered so far, things like transposing, inverting, etc., have been matrix-valued functions of the matrix in that the outcome of applying the operation to the matrix has yielded a matrix as the solution.

We have already seen an example of this in Section 1.3.7 where it was stated that the determinant of a 2×2 matrix is

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{3.0.1}$$

Similarly, it can be shown that the determinant of a scalar matrix is just the scalar itself, i.e. $|a| = a$, and the determinant of a 3×3 matrix is

$$\begin{aligned} |\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} \\ &\quad - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{31}a_{22}. \end{aligned} \tag{3.0.2}$$

Determinants occur throughout matrix algebra and its applications. They variously admit interpretations, *inter alia*, as volumes or areas in geometry and generalized variances in statistics. We have seen that they can occur in the formula for the inverse of a 2×2 matrix; more generally the formula for an inverse of arbitrary dimension can be expressed in terms of determinants. Our current interest in determinants stems from a particular application known as Cramer's Rule. Before discussing Cramer's Rule we will present two methods that will enable you to evaluate determinants. The first is a simple device that only applies to finding determinants of matrices of small dimension whereas the other method can be applied to find determinants of square matrices of arbitrary dimension.

3.1 Determinants of 2×2 and 3×3 Matrices

In Figure 3.1 we see a 2×2 matrix with two arrows passing through it; these arrows show how to combine the elements of this matrix to evaluate its determinant. The idea is to form products involving all the elements on a given arrow. There are

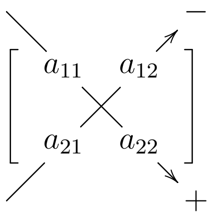


Figure 3.1: Evaluating the Determinant of a 2×2 Matrix

two arrows here and so we have two products, namely $a_{11}a_{22}$ and $a_{12}a_{21}$. The final final part of the trick is to sign these products and then add the signed products together. The way we sign the product is simply that we add the product coming from the arrow that slopes down as you move from left to right, that is the arrow with a negative slope, and subtract the product associated with the arrow that slope up us you move from left to right, that is the arrow with the positive slope. This rule suggests that, if $\mathbf{A}$ is a 2×2 matrix, $|\mathbf{A}| = +a_{11}a_{22} - a_{12}a_{21}$. Compare this result with that of Equation (3.0.1).

Example 3.1.1. The Determinant of a 2×2 Matrix.

$$\begin{vmatrix} 3 & 2 \\ 7 & 1 \end{vmatrix} = \begin{bmatrix} 3 & 2 \\ 7 & 1 \end{bmatrix} = +(3 \times 1) - (7 \times 2) = -11. \qquad \square$$

A similar device can be used to evaluate the determinant of a 3×3 matrix. Consider Figure 3.2 which presents a 3×3 array (we have omits the square brackets that would make the array a matrix because the diagram is already cluttered enough). The idea behind evaluating the determinant of a 3×3 matrix is similar to that used with 2×2 matrices. Again there are arrows that select combinations of elements of the matrix that should be multiplied together. Whereas for a 2×2 matrix the arrows led to products of pairs of elements, for 3×3 matrices we need to form products involving three elements.[1] Now the only diagonals that have three elements on them are the leading diagonal, running from (1,1) element to the (3,3) element, and the diagonal running from the (1,3) element to the (3,1) element of the matrix. Consequently, in order to pick up the required three elements the arrows need to bend, as illustrated in Figure 3.2. The basic idea is, however, the same as for determinants of 2×2 matrices. We add the products selected by downward sloping arrows (the solid lines in Figure 3.2) and subtract the products selected by downward sloping arrows (the dashed lines in Figure 3.2). If $\mathbf{A}$ is a 3×3 matrix then, by following the rule, we simply obtain Equation (3.0.2) as our expression for the determinant. [2]

[1] Although this device of using arrows does not extend to determinants of matrices with larger dimensions it is that case that determinants of 4×4 matrices involve products of four elements from the matrix, and the determinants of 5×5 matrices involve products of five elements from the matrix, and so on.

[2] What is given here is a variant of the mnemonic, named Sarrus' scheme, discovered by the French mathematician Pierre Frederic Sarrus (1798–1861). In his version, the scheme is illustrated

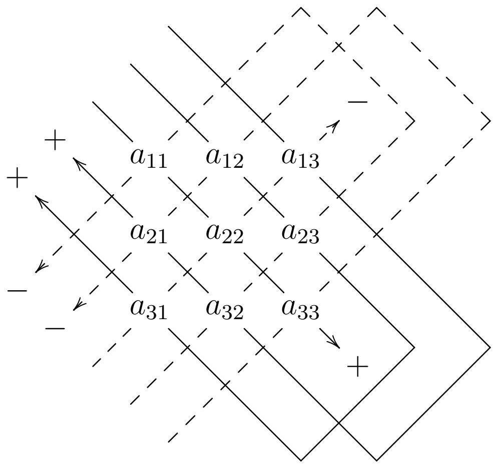


Figure 3.2: Evaluating the Determinant of a 3×3 Matrix

Example 3.1.2. The Determinant of a 3×3 Matrix.

$$\begin{aligned}\begin{vmatrix} 3 & 6 & 9 \\ 0 & 2 & 1 \\ 4 & 5 & 7 \end{vmatrix} &= +(3 \times 2 \times 7) + (6 \times 1 \times 4) + (9 \times 5 \times 0) \\ &\quad - (4 \times 2 \times 9) - (5 \times 1 \times 3) - (7 \times 6 \times 0) \\ &= -21. \end{aligned}$$

The appropriate 'arrow diagram' in this case is provided in Figure 3.3. □

Note that this device does not extend to the determinants of larger matrices. In order to evaluate determinants of larger matrices it is necessary to use the algorithm provided by Laplace's expansion that is discussed in the next section.

3.2 Properties of Determinants

If $\mathbf{X}$ and $\mathbf{Y}$ are $n \times n$ matrices, then:

(i) If any two rows (or columns) of $\mathbf{X}$ are interchanged then the sign of $\det(\mathbf{X})$ changes.

---

in terms of an augmented matrix, as follows

$$\begin{array}{ccccc} + & + & + & & \\ \multicolumn{5}{c}{\left|\begin{array}{ccc|cc} a_{11} & a_{12} & a_{13} & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} & a_{31} & a_{32} \end{array}\right.} \\ - & - & - & & \end{array}$$

Each arrow starting from the top row passes through three terms, the product of which areadded to the the value of the determinant, while the triplets associated with arrows starting in the bottom row are subtracted from the value of the determinant.

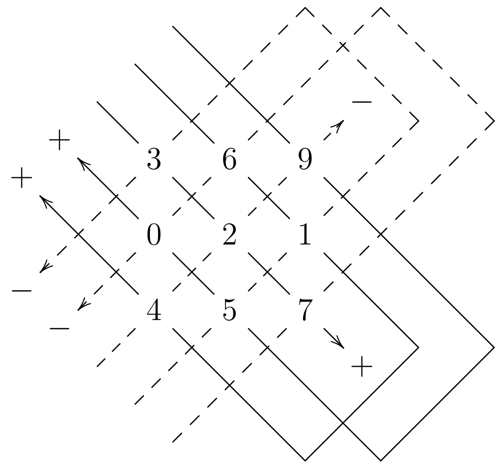


Figure 3.3: Arrow Diagram for Example 3.1.2

(ii) The determinant of $\mathbf{X}$ remains unchanged if any multiple of a row (or column) is added to any other row (or column).

(iii) Let $\mathbf{Y}$ be the matrix obtained when all the elements of a single row (or column) of $\mathbf{X}$ are multiplied by some constant k. Then $\det(\mathbf{Y}) = k\det(\mathbf{X})$.

(iv) If all of the entries in any row (or column) of $\mathbf{X}$ are zero then $\det(\mathbf{X}) = 0$.

(v) If any row (or column) of $\mathbf{X}$ can be written as a weighted sum of the other rows (or columns) of $\mathbf{X}$ then $\det(\mathbf{X}) = 0$. In particular, if any two rows (or columns) of $\mathbf{X}$ are either scalar multiples of each other or identical then $\det(\mathbf{X}) = 0$.

(vi) If $\mathbf{X}$ is triangular (or diagonal) then $\det(\mathbf{X}) = \displaystyle\prod_{i=1}^{n} x_{ii} = x_{11} \times x_{22} \times \ldots \times x_{nn}$.

(vii) If $\mathbf{Y} = k\mathbf{X}$ for some constant k, then $\det(\mathbf{Y}) = k^n \det(\mathbf{X})$.

(viii) $\det(\mathbf{XY}) = \det(\mathbf{X})\det(\mathbf{Y})$, provided that $\mathbf{X}$ and $\mathbf{Y}$ are both square.

(ix) $\det(\mathbf{X}^\top) = \det(\mathbf{X})$.

(x) $\det(\mathbf{X}^p) = (\det(\mathbf{X}))^p$. In particular, $\det(\mathbf{X}^{-1}) = \dfrac{1}{\det(\mathbf{X})}$.

Example 3.2.1. Elementary Row Operations and Determinants.

Using elementary row operations to triangulate the matrix, evaluate $\begin{vmatrix} 3 & 6 & 7 \\ 2 & 9 & 2 \\ 4 & 1 & 8 \end{vmatrix}$.

SOLUTION

$$\begin{vmatrix} 3 & 6 & 7 \\ 2 & 9 & 2 \\ 4 & 1 & 8 \end{vmatrix} \xrightarrow[R_3^*=R_3-\frac{4}{3}R_1]{R_2^*=R_2-\frac{2}{3}R_1} \begin{vmatrix} 3 & 6 & 7 \\ 0 & 5 & -8/3 \\ 0 & -7 & -4/3 \end{vmatrix} \xrightarrow[R_3^*=R_3+\frac{7}{5}R_2]{} \begin{vmatrix} 3 & 6 & 7 \\ 0 & 5 & -8/3 \\ 0 & 0 & -76/15 \end{vmatrix}$$

$$= 3 \times 5 \times -\frac{76}{15} = -76,$$

as before (see Example 3.3.3). □

Readers should convince themselves that the other properties hold by working some examples.

3.3 The Laplace Expansion of a Determinant

The Laplace expansion of a determinant can be applied to find the determinants of square matrices of arbitrary dimension. It is a recursive algorithm in the sense that it defines the determinant of a matrix in terms of determinants of matrices of smaller dimension. These latter determinants can be found in terms of determinants of matrices of yet smaller dimension. The process continues until the original determinant can be expressed in terms of determinants of scalar matrices, which are just the scalars themselves. Although this process will always work it can be extremely tedious and there can be considerable work reduction arising from judicious application of the expansion. The expansion itself is simple although there is a number of concepts to define. We shall do this by first defining the expansion and then defining ever more primitive concepts until everything is defined, in much the same way as the algorithm works.

Definition 3.3.1. Laplace Expansion of a Determinant.
Let $\mathbf{X}$ be a square matrix of order n then the *Laplace expansion of a determinant* can be written

$$\begin{aligned}
\det(\mathbf{X}) &= \sum_{j=1}^{n} x_{ij} c_{ij}, \text{ for any } i = 1, \ldots, n, && \text{(by rows)} \\
&= \sum_{i=1}^{n} x_{ij} c_{ij}, \text{ for any } j = 1, \ldots, n, && \text{(by columns)}
\end{aligned}$$

where c_{ij} denotes the cofactor of the ijth element of $\mathbf{X}$ and $\det(k) = k$ for any scalar k. □

In order to make this definition operational it is necessary to define the cofactors c_{ij}, which in turn require the definition of minors of a matrix.

Definition 3.3.2. Minors of a Matrix.
If $\mathbf{X}$ is a square matrix of order n, then the *minor* of the element x_{ij}, denoted $\mathbf{M}_{ij}$, is the determinant of the $(n-1) \times (n-1)$ matrix obtained by deleting the ith row and jth column from $\mathbf{X}$. $\mathbf{M}_{ii}$ $(i = 1, \ldots, n,)$ is called a *principal minor.* □

Example 3.3.1. Minors of a Matrix.
For the matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}, \quad M_{21} = \begin{vmatrix} x_{12} & x_{13} \\ x_{32} & x_{33} \end{vmatrix} \quad \text{and} \quad M_{33} = \begin{vmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{vmatrix}.$$

M_{33} is one of the three principal minors of $\mathbf{X}$, the other two being M_{11} and M_{22}. □

Definition 3.3.3. Cofactors of a Matrix.
The *cofactor* (c_{ij}) of the element x_{ij} is defined to be $c_{ij} = (-1)^{i+j} M_{ij}$. The *cofactor matrix* $\mathbf{C}$ is defined to be $\mathbf{C} = \{c_{ij}\}_{i,j=1,\ldots,n}$. □

Note that if $\mathbf{X}$ is $n \times n$ then so is $\mathbf{C}$.

Example 3.3.2. Cofactors.
Continuing the previous example:

$$c_{21} = (-1)^{2+1} M_{21} = -M_{21} \qquad \text{and} \qquad c_{33} = (-1)^{3+3} M_{33} = M_{33}.$$ □

Properties of Cofactors Cofactors have two important properties. The first is that provided in in Definition 3.3.1. The second is the following:

$$\begin{aligned} 0 &= \sum_{j=1}^{n} x_{ij} c_{kj}, \text{ for any } k \neq i = 1, \ldots, n, && \text{(by rows)} \\ &= \sum_{i=1}^{n} x_{ij} c_{ik}, \text{ for any } k \neq j = 1, \ldots, n. && \text{(by columns)} \end{aligned}$$

This last property follows because, by comparison with the Laplace expansion, the right-hand side is the expansion that you would obtain if you had replaced the kth row (or column) with the ith row (or column). Of course, that would mean that you have two identical rows (or columns) in $\mathbf{X}$ and so it must have a determinant of zero.

Cofactor matrices have the following additional properties:

(i) If $\rho(\mathbf{X}) = n$ then $\rho(\mathbf{C}) = n$ and $\det(\mathbf{X}) \neq 0$. A non-zero determinant is one definition of the non-singularity of a matrix.

(ii) If $\rho(\mathbf{X}) < n$ then $\det(\mathbf{X}) = 0$ and $\mathbf{X}$ is said to be singular. If $\rho(\mathbf{X}) = n - 1$ then $\rho(\mathbf{C}) = 1$, and if $\rho(\mathbf{X}) < n - 1$ then $\rho(\mathbf{C}) = 0$ ($\mathbf{C} = \mathbf{0}$).

The definition of cofactors was required for the definition of determinants. The fact that cofactors are defined in terms of determinants might lead one to believe that the the definition of a determinant is circular, although such a belief is mistaken. The essential feature of the definition of a cofactor is that it involves determinants of matrices of smaller order than the original matrix. Thus, the determinant of an $n \times n$ matrix involves cofactors which are determinants of $(n-1) \times (n-1)$ matrices, which in turn involve cofactors that are determinants of $(n-2) \times (n-2)$ matrices and so on. The process can continue in this way until one requires the determinants of scalars (1×1 matrices) which are just the scalars themselves. Before illustrating the calculation of a determinant in an example, one more concept shall be defined which, although not required in the definition of a determinant, is logically grouped with the definition of a cofactor.

Definition 3.3.4. Adjoint Matrix.
The *adjoint matrix*, or *adjugate matrix*, $\mathbf{\Psi}$ say, is defined to be the transpose of the cofactor matrix $\mathbf{C}$; that is, $\mathbf{\Psi} \equiv \operatorname{adj}(\mathbf{X}) = \mathbf{C}^\top$. □

Note that, for any square matrix $\mathbf{X}$ say, both the cofactor matrix and the adjoint matrix are well-defined, unlike the matrix inverse (say) which is only defined for non-singular matrices. Adjoint matrices have one very important property, that follows from the fundamental properties of cofactors; namely, for $\mathbf{X}$ $n \times n$,

$$\mathbf{X\Psi} = \det(\mathbf{X})\mathbf{I}_n. \tag{3.3.1}$$

That is, the terms on the leading diagonal of $\mathbf{X\Psi}$ correspond to a statement of the Laplace expansion and the off-diagonal elements capture the second property of cofactors.

Having defined cofactors it is now possible to calculate determinants.

Example 3.3.3. Laplace Expansion of a Determinant.

Find the determinants of (i) $\begin{bmatrix} 3 & 6 \\ 2 & 9 \end{bmatrix}$ and (ii) $\begin{bmatrix} 3 & 6 & 7 \\ 2 & 9 & 2 \\ 4 & 1 & 8 \end{bmatrix}$.

SOLUTION

(i) Expanding across the first row

$$\begin{vmatrix} 3 & 6 \\ 2 & 9 \end{vmatrix} = 3 \times (-1)^{1+1}|9| + 6 \times (-1)^{1+2}|2| = 27 - 12 = 15,$$

or, expanding down the second column

$$\begin{vmatrix} 3 & 6 \\ 2 & 9 \end{vmatrix} = 6 \times (-1)^{1+2}|2| + 9 \times (-1)^{2+2}|3| = -12 + 27 = 15.$$

(ii) Expanding across the third row

$$\begin{vmatrix} 3 & 6 & 7 \\ 2 & 9 & 2 \\ 4 & 1 & 8 \end{vmatrix} = 4 \times (-1)^{3+1} \begin{vmatrix} 6 & 7 \\ 9 & 2 \end{vmatrix} + 1 \times (-1)^{3+2} \begin{vmatrix} 3 & 7 \\ 2 & 2 \end{vmatrix} + 8 \times (-1)^{3+3} \begin{vmatrix} 3 & 6 \\ 2 & 9 \end{vmatrix}$$

$$= 4 \times (12 - 63) - 1 \times (6 - 14) + 8 \times (27 - 12) = -76.$$

□

3.4 The Diagonal Expansion of the Determinant of a Matrix

In the previous section we saw a version of the most commonly encountered technique for finding the determinant of a matrix. In this section we look at another method that will prove useful in our study of eigenvalues on Chapter 6. The idea behind this approach comes from the fact that any square matrix, $\mathbf{A}^*$ say, can be expressed as the sum of two other matrices, $\mathbf{A}+\mathbf{D}$ say, where $\mathbf{D}$ denotes a diagonal matrix and $\mathbf{A} = \mathbf{A}^* - \mathbf{D}$. The diagonal expansion is a simple way of obtaining $\det(\mathbf{A} + \mathbf{D})$ as a polynomial in the elements of $\mathbf{D}$ which, as suggested above, is sometimes of interest to us. It is important to note that no matter how one chooses to evaluate a determinant its value remains unchanged. These different expansion metods are of interest only for the mathematical convenience they may afford under certain circumstances. Let us first illustrate the idea in the case of a 2×2 matrix.

Example 3.4.1. Diagonal Expansion of the Determinant of a 2×2 Matrix. We know that

$$\det(\mathbf{A}^*) = \begin{vmatrix} a_{11}^* & a_{12} \\ a_{21} & a_{22}^* \end{vmatrix} = a_{11}^* a_{22}^* - a_{12}a_{21}.$$

Suppose that we replaced a_{11}^* by $a_{11} + d_1$ and a_{22}^* by $a_{22} + d_2$. That is, we wrote $\mathbf{A}^* = \mathbf{A} + \mathbf{D}$, where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \text{and} \qquad \mathbf{D} = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}.$$

Then, we might write

$$\begin{aligned} \det(\mathbf{A}^*) &= (a_{11} + d_1)(a_{22} + d_2) - a_{12}a_{21} = d_1d_2 + d_1a_{22} + d_2a_{11} + a_{11}a_{22} - a_{12}a_{21} \\ &= d_1d_2 + d_1a_{22} + d_2a_{11} + \det(\mathbf{A}), \end{aligned} \tag{3.4.1}$$

which is a polynomial in the elements of $\mathbf{D}$. □

Now let's try the same thing with a 3×3 matrix.

Example 3.4.2. Diagonal Expansion of the Determinant of a 2×2 Matrix. Writing $\mathbf{A}^* = \mathbf{A} + \mathbf{D}$, where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad \text{and} \qquad \mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix},$$

we can then show that

$$\begin{aligned} \det(\mathbf{A}^*) &= d_1d_2d_3 + d_1d_2a_{33} + d_1d_3a_{22} + d_2d_3a_{11} \\ &\quad + d_1 \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + d_2 \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} + d_3 \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} + \det(\mathbf{A}). \end{aligned} \tag{3.4.2}$$

which is a polynomial in the elements of $\mathbf{D}$. □

The pattern that is important to recognise as you look at (3.4.1) and (3.4.2) is that the coefficients on the various products of the elements of $\mathbf{D}$ are the corresponding principal minors of $\mathbf{A}$ that you obtain when deleting from $\mathbf{A}$ those rows and columns corresponding to the subscripts on the elements of $\mathbf{D}$ in the product. For example, in each case, the coefficient on the term d_1 is the principal minor of $\mathbf{A}$ corresponding to the term a_{11}. This is sometimes called the *complementary principal minor in* $\mathbf{A}$. In (3.4.2), the coefficient on the term d_1d_3 is the principal minor obtained if one deletes the first and third rows and columns from $\mathbf{A}$, namely a_{22}.

Now, our interest in diagonal expansons will be restricted to the case where $d_1 = \cdots = d_n = d$ say. In this case, (3.4.1) becomes

$$\det(\mathbf{A}^*) = d^2 + d(a_{11} + a_{22}) + \det(\mathbf{A}) = d^2 + d\,\mathrm{tr}(\mathbf{A}) + \det(\mathbf{A}) \tag{3.4.3}$$

and (3.4.2) becomes

$$\det(\mathbf{A}^*) = d^3 + d^2\,(a_{11} + a_{22} + a_{33})$$

$$
\begin{aligned}
&+ d\left(\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}\right) + \det(\mathbf{A}) \\
= {} & d^3 + d^2 \operatorname{tr}(\mathbf{A}) \\
&+ d\left(\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}\right) + \det(\mathbf{A}).
\end{aligned} \tag{3.4.4}
$$

This would seem like a reasonably efficient expression if we had some notation for that sum of determinants of 2×2 matrices. It should come as no surprise that we do.

If one takes a close look at (3.4.3) and (3.4.4), then the following patterns emerge. First, if $\mathbf{A}^*$ is an $n \times n$ matrix then the overall expression is a polynomial of order n in d. Second, the coefficient on each term d^k is a polynomial in the elements of $\mathbf{A}$ of order $n-k$. Thus, the coefficient on the term d^n is a polynomial of order $n - n = 0$ in the elements of $\mathbf{A}$. That is, the coefficient is simply the constant 1. Similarly, the coefficient on the term d^{n-1} is a polynomial in the elements of $\mathbf{A}$ of order $n-(n-1) = 1$. In each case it is just the sum of terms on the leading diagonal of $\mathbf{A}$, that is the trace of $\mathbf{A}$. In (3.4.4) we see that the coefficient on the d terms is a sum of second order polynomials in the elements of $\mathbf{A}$ which correponds to the set of leading principal minors of $\mathbf{A}$. Indeed, if you think about it, in the case of a 3×3 matrix, the trace of the matrix is also a sum of principal minors, except that they are constructed with respect to two elements on the leading diagonal rather than just one. Consequently, a notation has developed which allows us to write, for an $n \times n$ matrix $\mathbf{A}^* = \mathbf{A} + \mathbf{D}$, with $\mathbf{D} = d\mathbf{I}_n$, the

$$
\det(\mathbf{A}^*) = \sum_{j=0} d^{n-j} \operatorname{tr}_j(\mathbf{A}), \tag{3.4.5}
$$

where

$$
\begin{aligned}
\operatorname{tr}_0(\mathbf{A}) &= 1, \\
\operatorname{tr}_1(\mathbf{A}) &= \operatorname{tr}(\mathbf{A}), \\
&\ \vdots \\
\operatorname{tr}_k(\mathbf{A}) &= \sum_{\kappa} M_\kappa, \\
&\ \vdots \\
\operatorname{tr}_n(\mathbf{A}) &= \det(\mathbf{A}),
\end{aligned} \tag{3.4.6}
$$

where the notation $\sum_\kappa M_\kappa$ should be read as meaning the sum of all possible principal minors of order k. Note that in general we would expect each such sum to contain $\binom{n}{k}$ such terms. For example, in (3.4.4) where $n = 3$, the coefficient on the d should be a polynomial of order $k = 2$ in the elements of $\mathbf{A}$. Thus, we should expect this polynomial to be comprised of

$$
\binom{3}{2} = \frac{3!}{2!\,(3-2)!} = \frac{3 \times 2!}{2! \times 1!} = 3
$$

principal minors, as the ordering of the subscripts doesn't matter,which is indeed the case. If we had a 4×4 matrix initially then the coefficient on d^2 would be a sum of 6 principle minors of order 2, which covers all the different ways of deleting two rows and two columns at a time. That is, the pairs of rows and columns deleted would be (1,2), (1,3), (1,4), (2,3), (2,4), (3,4), which is the complete set of possibilities.

3.5 Matrix Inversion and Determinants

It can be shown that if $\mathbf{X}$ is a non-singular matrix then[3] $\mathbf{X}^{-1} = \dfrac{1}{\det(\mathbf{X})} \operatorname{adj}(\mathbf{X})$.

Example 3.5.1. Matrix Inverse and Determinants.
If

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix},$$

then the cofactor matrix is

$$\mathbf{C} = \begin{bmatrix} x_{22} & -x_{21} \\ -x_{12} & x_{11} \end{bmatrix}$$

and the adjoint matrix is

$$\mathbf{\Psi} = \operatorname{adj}(\mathbf{X}) = \mathbf{C}^\top = \begin{bmatrix} x_{22} & -x_{12} \\ -x_{21} & x_{11} \end{bmatrix}.$$

Next, $\det(\mathbf{X}) = x_{11}x_{22} - x_{12}x_{21}$. Combining these results yields

$$\mathbf{X}^{-1} = \frac{1}{x_{11}x_{22} - x_{12}x_{21}} \begin{bmatrix} x_{22} & -x_{12} \\ -x_{21} & x_{11} \end{bmatrix}.$$

If $\det(\mathbf{X}) = 0$ then $\mathbf{X}^{-1}$ is not defined because of the division by zero. □

In general, any matrix with zero determinant is singular (or non-invertible), meaning that the matrix inverse is not defined, because of the division by zero observed in this example.

Example 3.5.2. Inverting a 3×3 Matrix Using Determinants.
In Example 3.2.1 it was shown that

$$\det \mathbf{X} = \begin{vmatrix} 3 & 6 & 7 \\ 2 & 9 & 2 \\ 4 & 1 & 8 \end{vmatrix} = -76.$$

The cofactor matrix of $\mathbf{X}$ can be evaluated as

$$\mathbf{C} = \begin{bmatrix} \begin{vmatrix} 9 & 2 \\ 1 & 8 \end{vmatrix} & -\begin{vmatrix} 2 & 2 \\ 4 & 8 \end{vmatrix} & \begin{vmatrix} 2 & 9 \\ 4 & 1 \end{vmatrix} \\ \\ -\begin{vmatrix} 6 & 7 \\ 1 & 8 \end{vmatrix} & \begin{vmatrix} 3 & 7 \\ 4 & 8 \end{vmatrix} & -\begin{vmatrix} 3 & 6 \\ 4 & 1 \end{vmatrix} \\ \\ \begin{vmatrix} 6 & 7 \\ 9 & 2 \end{vmatrix} & -\begin{vmatrix} 3 & 7 \\ 2 & 2 \end{vmatrix} & \begin{vmatrix} 3 & 6 \\ 2 & 9 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} 70 & -8 & -34 \\ -41 & -4 & 21 \\ -51 & 8 & 15 \end{bmatrix}.$$

[3]This is simply a re-arrangement of (3.3.1).

By definition $\mathbf{X}^{-1} = |\mathbf{X}|^{-1}\mathbf{C}^\top$. That is,

$$\mathbf{X}^{-1} = \frac{1}{76}\begin{bmatrix} 70 & -41 & -51 \\ -8 & -4 & 8 \\ -34 & 21 & 15 \end{bmatrix}.$$

Checking that $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}_3$ is left as an exercise. □

3.6 Cramer's Rule

Our interest in matrices has largely been driven by a desire to solve systems of equations. Sometimes we are only interested in the values taken by subset of the unknowns in a system of equations. Cramer's Rule is a method for determining the value of a single unknown in a system of equations with an arbitrary numbers of unknowns.

Definition 3.6.1. Cramer's Rule.
Consider the system of equations $\mathbf{\Gamma}\mathbf{x} = \mathbf{b}$, where $\mathbf{\Gamma}$ is an $n \times n$ matrix of known constants, $\mathbf{b}$ is an $n \times 1$ vector of known constants and $\mathbf{x}$ is an $n \times 1$ vector of unknowns. If we represent that n columns of $\mathbf{\Gamma}$ by $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$, so that $\mathbf{\Gamma} = [\boldsymbol{\gamma}_1 \ \ldots \ \boldsymbol{\gamma}_n]$, then x_i, the ith element of $\mathbf{x}$, can be obtained by evaluating

$$x_i = \frac{|\mathbf{\Gamma}_i^*|}{|\mathbf{\Gamma}|},$$

where $\mathbf{\Gamma}_i^*$ is the matrix obtained when the ith column of $\mathbf{\Gamma}$ is replaced by $\mathbf{b}$.

Example 3.6.1. Using Cramer's Rule with 2×2 Matrices.
In Example 1.6.1 have seen that the solution to the system of equations

$$\begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 11 \\ 22 \end{bmatrix}$$

was $x_1 = 5$ and $x_2 = 1$. We could have used Cramer's Rule to solve for either x_1 or x_2 on their own without solving for the other one. In terms of Definition 3.6.1,

$$\mathbf{\Gamma} = \begin{bmatrix} 2 & 1 \\ 3 & 7 \end{bmatrix} \qquad \text{and} \qquad \mathbf{b} = \begin{bmatrix} 11 \\ 22 \end{bmatrix}.$$

If we wish to solve for x_1 alone then

$$\mathbf{\Gamma}_1^* = \begin{bmatrix} 11 & 1 \\ 22 & 7 \end{bmatrix}.$$

Thus,

$$x_1 = \frac{\begin{vmatrix} 11 & 1 \\ 22 & 7 \end{vmatrix}}{\begin{vmatrix} 2 & 1 \\ 3 & 7 \end{vmatrix}} = \frac{11 \times 7 - 1 \times 22}{2 \times 7 - 1 \times 3} = \frac{55}{11} = 5,$$

as before. Similarly, if we had wished to solve for x_2 alone then

$$\mathbf{\Gamma}_2^* = \begin{bmatrix} 2 & 11 \\ 3 & 22 \end{bmatrix}.$$

Thus,

$$x_2 = \frac{\begin{vmatrix} 2 & 11 \\ 3 & 22 \end{vmatrix}}{\begin{vmatrix} 2 & 1 \\ 3 & 7 \end{vmatrix}} = \frac{2 \times 22 - 11 \times 3}{2 \times 7 - 1 \times 3} = \frac{11}{11} = 1,$$

as before. Note that the denominator of the calculation is the same for both x_1 and x_2 and so need only be calculated once if you are solving for more than one variable in a system. □

Example 3.6.2. Using Cramer's Rule with 3×3 Matrices.
Use Cramer's Rule to solve for x_2 in the system of equations

$$\begin{aligned} x_1 + \ x_2 + \ x_3 &= 2 \\ 2x_1 - \ x_2 + 3x_3 &= 0 \\ x_1 + 2x_2 - \ x_3 &= 5. \end{aligned}$$

SOLUTION

First write the system in matrix notation

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & 3 \\ 1 & 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 5 \end{bmatrix}.$$

By Cramer's Rule,

$$x_2 = \frac{\begin{vmatrix} 1 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 5 & -1 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ 2 & -1 & 3 \\ 1 & 2 & -1 \end{vmatrix}} = \frac{5}{5} = 1.$$

It would be a useful exercise to use Cramer's Rule to also show that $x_1 = 2$ and $x_3 = -1$. □

Example 3.6.3. Another Example of Cramer's Rule with 3×3 Matrices.
Use Cramer's rule to solve for z in the system:

$$\begin{aligned} 2x - y + 3z &= 12 \\ x + y - z &= -3 \\ x + 2y - 3z &= -10 \end{aligned}$$

SOLUTION

In matrix notation we have

$$\begin{bmatrix} 2 & -1 & 3 \\ 1 & 1 & -1 \\ 1 & 2 & -3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 12 \\ -3 \\ -10 \end{bmatrix}.$$

By Cramer's rule

$$z = \frac{\begin{vmatrix} 2 & -1 & 12 \\ 1 & 1 & -3 \\ 1 & 2 & -10 \end{vmatrix}}{\begin{vmatrix} 2 & -1 & 3 \\ 1 & 1 & -1 \\ 1 & 2 & -3 \end{vmatrix}}.$$

Expanding the numerator down the third column

$$\begin{aligned} \begin{vmatrix} 2 & -1 & 12 \\ 1 & 1 & -3 \\ 1 & 2 & -10 \end{vmatrix} &= 12 \times (-1)^{1+3} \begin{vmatrix} 1 & 1 \\ 1 & 2 \end{vmatrix} - 3 \times (-1)^{2+3} \begin{vmatrix} 2 & -1 \\ 1 & 2 \end{vmatrix} - 10 \times (-1)^{3+3} \begin{vmatrix} 2 & -1 \\ 1 & 1 \end{vmatrix} \\ &= 12 \times 1 + 3 \times 5 - 10 \times 3 \\ &= -3. \end{aligned}$$

Similarly, expanding the denominator down the third column,

$$\begin{vmatrix} 2 & -1 & 3 \\ 1 & 1 & -1 \\ 1 & 2 & -3 \end{vmatrix} = 3 \times 1 + 1 \times 5 - 3 \times 3 = -1.$$

Consequently,

$$z = \frac{-3}{-1} = 3.$$

Note: I have expanded both numerator and denominator down the third column to use the same cofactors, thereby reducing calculations. □

As further exercises readers could use this same example to establish that $x = 1$ and $y = -1$.

3.6.1 A Formal Development of Cramer's Rule

The reasoning underlying Cramer's Rule is as follows. Letting c_{ij} denote the cofactor of γ_{ij} we have, by definition,

$$\begin{aligned} \mathbf{X} = \mathbf{\Gamma}^{-1}\mathbf{B} &= \frac{1}{\det(\mathbf{\Gamma})} (\operatorname{adj}(\mathbf{\Gamma}))\, \mathbf{B} \\ &= \frac{1}{\det(\mathbf{\Gamma})} \begin{bmatrix} c_{11} & \cdots & c_{n1} \\ \vdots & \ddots & \vdots \\ c_{1n} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \frac{1}{\det(\mathbf{\Gamma})} \begin{bmatrix} \sum_{i=1}^{n} b_i c_{i1} \\ \vdots \\ \sum_{i=1}^{n} b_i c_{in} \end{bmatrix}. \end{aligned}$$

Each sum can be interpreted as the Laplace expansion of a determinant down the jth column of some matrix, where $\mathbf{B}$ is the jth column of that matrix. In particular,

$$\mathbf{X} = \frac{1}{\det(\mathbf{\Gamma})} \begin{bmatrix} \det(\mathbf{\Delta}_1) \\ \vdots \\ \det(\mathbf{\Delta}_n) \end{bmatrix},$$

where $\mathbf{\Delta}_j$ is the matrix obtained when the jth column of $\mathbf{\Gamma}$ is replaced by $\mathbf{B}$. For example,

$$\mathbf{\Delta}_1 = \begin{bmatrix} b_1 & \gamma_{12} & \cdots & \gamma_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ b_n & \gamma_{n2} & \cdots & \gamma_{nn} \end{bmatrix}, \quad \mathbf{\Delta}_2 = \begin{bmatrix} \gamma_{11} & b_1 & \gamma_{13} & \cdots & \gamma_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{n1} & b_n & \gamma_{n3} & \cdots & \gamma_{nn} \end{bmatrix}, \quad \text{etc.}$$

That is,

$$x_1 = \frac{\det(\mathbf{\Delta}_1)}{\det(\mathbf{\Gamma})}, \qquad x_2 = \frac{\det(\mathbf{\Delta}_2)}{\det(\mathbf{\Gamma})}, \qquad \text{etc.}$$

3.7 Quadratic Forms and Definiteness

In previous sections we have explored how the algebraic ideas that we take for granted when dealing with numbers can be extended to matrix algebra. In this section we seek to extend another very simple idea to its matrix analogue. The idea is that of the sign of a number. When one thinks about the problem it is not at all clear how to proceed, the first problem being to formally specify the problem. When dealing with numbers, the sign is a feature of a single entity, namely the number. A matrix potentially has many elements and so the problem can be thought of as trying to find some summary measure of the characteristics of all of those elements that might fruitfully be thought of as having a property which is analogous to the sign of a number. The convention that we will adopt is based on the notion of a quadratic form. This is particularly convenient because quadratic forms arise naturally in both economics and statistics in circumstances where the sign of the quadratic form is of interest.

Definition 3.7.1. Quadratic Form.
Let $\mathbf{z}$ be an $n \times 1$ vector and $\mathbf{A}$ be an $n \times n$ (symmetric) matrix then the expression $Q(\mathbf{A}, \mathbf{z}) = \mathbf{z}^\top \mathbf{A} \mathbf{z}$ is termed a *quadratic form*.[4] All quadratic forms have the alternative representation

$$Q(\mathbf{A}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} z_i z_j. \qquad \square$$

Example 3.7.1. Sums of Squares Deviations as a Quadratic Form.
A sample variance can be defined as

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2, \qquad \text{where} \quad \overline{x} = \sum_{i=1}^{N} \frac{x_i}{N}.$$

Defining the symmetric matrix $\mathbf{M}_\imath = \mathbf{I}_N - \imath(\imath^\top \imath)^{-1} \imath^\top = \mathbf{I}_N - N^{-1} \imath\imath^\top$, where $\imath = [1, \ldots, 1]^\top$ is an $N \times 1$ vector consisting entirely of ones, the sum of squared deviations about the sample mean can be written as $\sum_{i=1}^{N} (x_i - \overline{x})^2 = \mathbf{x}^\top \mathbf{M}_\imath \mathbf{x}$, where $\mathbf{x}^\top = [x_1, \ldots, x_N]$. $\square$

[4] The symmetry of $\mathbf{A}$ is not essential to the definition of $Q(\mathbf{A}, \mathbf{z})$ but may as well be assumed because $Q(\mathbf{A}, \mathbf{z}) = Q(\mathbf{B}, \mathbf{z})$, where $\mathbf{B} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)$ is a symmetric matrix.

Definition 3.7.2. Definiteness of a Matrix.
If $Q(\mathbf{A}, \mathbf{z}) > 0$ for all $\mathbf{z} \neq \mathbf{0}$ then $\mathbf{A}$ is said to be *positive definite*. If $Q(\mathbf{A}, \mathbf{z}) < 0$ for all $\mathbf{z} \neq \mathbf{0}$ then $\mathbf{A}$ is said to be *negative definite*. If $Q(\mathbf{A}, \mathbf{z}) \geq 0$ for all $\mathbf{z} \neq \mathbf{0}$ then $\mathbf{A}$ is said to be *positive semi-definite* or *non-negative definite*. If $Q(\mathbf{A}, \mathbf{z}) \leq 0$ for all $\mathbf{z} \neq \mathbf{0}$ then $\mathbf{A}$ is said to be *negative semi-definite* or *non-positive definite*. If none of these relations can be established then the matrix $\mathbf{A}$ is said to indefinite. □

In Section 3.3 the idea of a principal minor was introduced. A similar idea can be defined along the following lines:

Definition 3.7.3. Leading Principal Minors of a Matrix.
The k^{th} order *leading principal minor* of the square matrix $\mathbf{A}$ is the determinant of the $k \times k$ matrix consisting of the first k rows and columns of $\mathbf{A}$.

The definition of leading principal minors provides an equivalent definition of the definiteness of a matrix that is often easier to establish than the original definition.

Definition 3.7.4. Definiteness and Minors. (i) The quadratic form $Q(\mathbf{A}, \mathbf{z})$ is positive definite if and only if all leading principal minors of $\mathbf{A}$ are positive.

(ii) The quadratic form $Q(\mathbf{A}, \mathbf{z})$ is negative definite if and only if the k^{th} order leading principal minors of the $n \times n$ matrix $\mathbf{A}$ have sign $(-1)^k$, $k = 1, \ldots, n$. □

Consider the symmetric matrix

$$\mathbf{H} = \begin{bmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{bmatrix}.$$

Definiteness requires that $|\mathbf{H}| = x_{11}x_{22} - x_{12}^2 > 0$. Should $|\mathbf{H}| \leq 0$ then $\mathbf{H}$ is an indefinite matrix. If $|\mathbf{H}| > 0$ then the only thing that need be examined is the sign of x_{11}. Note that because $-x_{12}^2 \leq 0$, definiteness in the 2×2 case requires that x_{11} and x_{22} have the same sign. This is a fairly simple set of conditions to check.

Example 3.7.2. Definiteness of a 2×2 Matrix.
Check the definiteness of the matrices

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 7 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} -1 & 3 \\ 3 & 7 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} -3 & 1 \\ 1 & -7 \end{bmatrix}.$$

$\mathbf{A}$ has $|\mathbf{A}| = 20 > 0$ and $a_{11} = 3 > 0$ and so it is positive definite. To see that this must be true observe that

$$Q(\mathbf{A}, \mathbf{z}) = 3z_1^2 + 7z_2^2 + 2z_1z_2 = 2z_1^2 + 6z_2^2 + (z_1 + z_2)^2 > 0, \quad \text{provided } \mathbf{z} \neq 0.$$

$\mathbf{B}$ is indefinite as b_{11} and b_{22} have different signs and $\mathbf{C}$ is negative definite as $|\mathbf{C}| = 20 > 0$ and $c_{11} = -3 < 0$. □

3.8 The Determinant of a Vandermonde Matrix

The Vandermonde matrix was introduced in Section 1.2. Its special structure allows us to derive a general result about its determinant, which is of importance in polynomial interpolation. To begin, we can sweep out the first column by a series of elementary row operations in which we subtract the first row from the remaining rows. We note that an elementary row operation that leaves the determinant unchanged (see Property (ii)). Hence, from (1.2.1)

$$|V_n| = \begin{vmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-2} & x_1^{n-1} \\ 0 & x_2 - x_1 & x_2^2 - x_1^2 & \dots & x_2^{n-2} - x_1^{n-2} & x_2^{n-1} - x_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & x_n - x_1 & x_n^2 - x_1^2 & \dots & x_n^{n-2} - x_1^{n-2} & x_n^{n-1} - x_1^{n-1} \end{vmatrix}.$$

Our second step is to sweep out the first row by a series of elementary column operations where, first, we subtract x_1 times the $(n-1)$th from the nth column. We repeat this procedure by subtracting x_1 times the $(n-2)$nd column from the $(n-1)$th column, and so on until we subtract x_1 times the 1st column from the 2nd column. This yields

$$|V_n| = \begin{vmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & x_2 - x_1 & (x_2 - x_1)x_2 & \dots & (x_2 - x_1)x_2^{n-3} & (x_2 - x_1)x_2^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & x_n - x_1 & (x_n - x_1)x_n & \dots & (x_n - x_1)x_n^{n-3} & (x_n - x_1)x_n^{n-2} \end{vmatrix}$$

and, again, the determinant remains unchanged as a consequence of Property (ii). Observe that the elements of each row share a common factor, i.e. in the ith row the common factor is $x_i - x_1$. By Property (iii), we see that

$$\begin{aligned} |V_n| &= (x_2 - x_1) \begin{vmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & x_2 & \dots & x_2^{n-3} & x_2^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & x_n - x_1 & (x_n - x_1)x_n & \dots & (x_n - x_1)x_n^{n-3} & (x_n - x_1)x_n^{n-2} \end{vmatrix} \\ &= \prod_{i=2}^{n} (x_i - x_1) \begin{vmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & x_2 & \dots & x_2^{n-3} & x_2^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & x_n & \dots & x_n^{n-3} & x_n^{n-2} \end{vmatrix} \\ &= \prod_{i=2}^{n} (x_i - x_1) |V_{n-1}| = \prod_{i=2}^{n} (x_i - x_1) \prod_{j=3}^{n} (x_j - x_2) |V_{n-2}| \end{aligned}$$

and so on. This proceeds until the final determinant to be resolved is

$$|V_2| = \begin{vmatrix} 1 & x_{n-1} \\ 1 & x_n \end{vmatrix} = x_n - x_{n-1}.$$

We might lay out all the terms of the final product in a triangular array:

$$\begin{array}{cccccc} (x_2 - x_1) & (x_3 - x_1) & \dots & (x_{n-1} - x_1) & (x_n - x_1) \\ & (x_3 - x_2) & \dots & (x_{n-1} - x_2) & (x_n - x_2) \\ & & \ddots & \vdots & \vdots \\ & & & (x_{n-1} - x_{n-2}) & (x_n - x_{n-2}) \\ & & & & (x_n - x_{n-1}) \end{array}$$

This product can be represented in different ways. Equally valid are

$$|V_n| = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} (x_j - x_i) \tag{3.8.1}$$

and

$$|V_n| = \prod_{1 \leq i < j \leq n} (x_j - x_i).$$

Example 3.8.1. Determinant of $|V_3|$.

By (3.8.1)

$$|V_3| = \prod_{i=1}^{2} \prod_{j=i+1}^{3} (x_j - x_i) = (x_2 - x_1)(x_3 - x_1)(x_3 - x_2). \qquad \square$$

Chapter 4

Partitioned Matrices

4.1 Definition and Elementary Operations

Consider an $m \times n$ matrix $\mathbf{A}$, which is a rectangular array of real numbers

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

For integer p and q, such that $1 \le p \le m$ and $1 \le q \le n$, we can define

$$\underset{p\times q}{\mathbf{A}_{11}} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,q} \\ \vdots & \ddots & \vdots \\ a_{p,1} & \cdots & a_{p,q} \end{bmatrix} \qquad \underset{p\times(n-q)}{\mathbf{A}_{12}} = \begin{bmatrix} a_{1,q+1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{p,q+1} & \cdots & a_{p,n} \end{bmatrix}$$

$$\underset{(m-p)\times q}{\mathbf{A}_{21}} = \begin{bmatrix} a_{p+1,1} & \cdots & a_{p+1,q} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,q} \end{bmatrix} \qquad \underset{(m-p)\times(n-q)}{\mathbf{A}_{22}} = \begin{bmatrix} a_{p+1,q+1} & \cdots & a_{p+1,n} \\ \vdots & \ddots & \vdots \\ a_{m,q+1} & \cdots & a_{m,n} \end{bmatrix}$$

so that

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

We say that $\mathbf{A}$ is a *partitioned* matrix.

Partitioned matrices are, at the end of the day, just matrices and so satisfy all of the usual properties of matrices. For example, let

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$$

have the same dimensions as $\mathbf{A}$ and its sub-matrices $\mathbf{B}_{11}$, $\mathbf{B}_{12}$, $\mathbf{B}_{21}$, and $\mathbf{B}_{22}$ have the same dimensions as the corresponding sub-matrices of $\mathbf{A}$. Then $\mathbf{B}$ is said to be *partitioned conformably under matrix addition* with $\mathbf{A}$ and

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} \\ \mathbf{A}_{21} + \mathbf{B}_{21} & \mathbf{A}_{22} + \mathbf{B}_{22} \end{bmatrix} = \mathbf{B} + \mathbf{A}$$

because addition is commutative for conformable matrices. Matrix multiplication requires a little more. Specifically, given our definition of $\mathbf{A}$, if $\mathbf{B}_{11}$ is of dimension $r \times s$, $\mathbf{B}_{12}$ is $r \times v$, $\mathbf{B}_{21}$ is $t \times s$, and $\mathbf{B}_{22}$ is $t \times v$ then

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}$$

if and only if $q = r$ and $n = q + t$. Similarly,

$$\mathbf{BA} = \begin{bmatrix} \mathbf{B}_{11}\mathbf{A}_{11} + \mathbf{B}_{12}\mathbf{A}_{21} & \mathbf{B}_{11}\mathbf{A}_{12} + \mathbf{B}_{12}\mathbf{A}_{22} \\ \mathbf{B}_{21}\mathbf{A}_{11} + \mathbf{B}_{22}\mathbf{A}_{21} & \mathbf{B}_{21}\mathbf{A}_{12} + \mathbf{B}_{22}\mathbf{A}_{22} \end{bmatrix}$$

requires $s = p$ and $m = v + p$. If either of these products can be formed in this way then we would say that the partitioned matrices are *conformable under matrix multiplication*. Note that one or other of the products $\mathbf{AB}$ or $\mathbf{BA}$ might be available but they will only both be available if, in addition to the restrictions given above, we also have $n = m$, i.e., $\mathbf{A}$ and $\mathbf{B}$ are square matrices of the same dimension which implies that $\mathbf{A}_{11}$ and $\mathbf{B}_{11}$ have the same dimensions, as do $\mathbf{A}_{22}$ and $\mathbf{B}_{22}$. We will, hereafter, assume this to be the case unless indicated otherwise. Specifically, we will assume that $\mathbf{A}$ is $n \times n$ and that $\mathbf{A}_{11}$ is $p \times p$.

An important device when working with partitioned matrices is the following pair of decompositions:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0}_{p,n-p} \\ \mathbf{A}_{21} & \mathbf{I}_{n-p} \end{bmatrix} \begin{bmatrix} \mathbf{I}_p & \mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{0}_{n-p,p} & \mathbf{A}_{22\cdot 1} \end{bmatrix} \tag{4.1.1a}$$

$$= \begin{bmatrix} \mathbf{I}_p & \mathbf{A}_{12} \\ \mathbf{0}_{n-p,p} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11\cdot 2} & \mathbf{0}_{p,n-p} \\ \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I}_{n-p} \end{bmatrix}, \tag{4.1.1b}$$

where

$$\mathbf{A}_{11\cdot 2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \tag{4.1.2a}$$

and

$$\mathbf{A}_{22\cdot 1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}. \tag{4.1.2b}$$

Two things to note about these decompositions: (i) (4.1.1a) does not require $\mathbf{A}_{22}$ to be non-singular and (4.1.1b) does not require $\mathbf{A}_{11}$ to be non-singular, and (ii) a useful aide-mémoire is that subscripts should match, e.g., $\mathbf{A}_{12}$ can pre-multiply $\mathbf{A}_{22}^{-1}$ but not $\mathbf{A}_{11}^{-1}$ whereas $\mathbf{A}_{12}$ can post-multiply $\mathbf{A}_{11}$ but not $\mathbf{A}_{22}$, and so on.

4.2 Determinant of a Partitioned Matrix

As a starting point, consider the problem of finding the determinant of the $m \times m$ matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{I}_p & \mathbf{X} \\ \mathbf{0}_{m \times p} & \mathbf{Z} \end{bmatrix}.$$

If we expand down the first column then we see that

$$\det \mathbf{W} = \det \begin{bmatrix} \mathbf{I}_{p-1} & \mathbf{X}_1 \\ \mathbf{0}_{m \times (p-1)} & \mathbf{Z} \end{bmatrix}$$

where $\mathbf{X}_1$ denotes what is left of $\mathbf{X}$ after the first row is deleted. All other terms in the expansion are identically equal to zero because all other terms in the first column are zero. If we iterate on this procedure (expanding down the first column of the matrix) then, after p iterations, we see that $\det \mathbf{W} = \det \mathbf{Z}$. Of particular interest is the fact that $\mathbf{X}$ is of no relevance in determining the value of the determinant. If we had been asked to find the value of $\mathbf{W}'$, we could proceeded by iteratively expanding across the first row of the determinant of interest to obtain $\det \mathbf{W}' = \det \mathbf{Z}'$. However, expanding $\det \mathbf{Z}$ down the first column, say, must yield identical results to expanding $\det \mathbf{Z}'$ across the first row (prove it!) and so we are drawn to the conclusion that $\det \mathbf{Z} = \det \mathbf{Z}'$ and hence $\det \mathbf{W} = \det \mathbf{W}'$.

Using these results, together with the observation that if $\mathbf{A}$ and $\mathbf{B}$ both $n \times n$ then $\det(\mathbf{AB}) = \det \mathbf{A} \times \det \mathbf{B}$, we see from (4.1.1) that

$$\det \mathbf{A} = \det \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0}_{p,(n-p)} \\ \mathbf{A}_{21} & \mathbf{I}_{n-p} \end{bmatrix} \times \det \begin{bmatrix} \mathbf{I}_p & \mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{0}_{n-p,p} & \mathbf{A}_{22\cdot 1} \end{bmatrix} = \det \mathbf{A}_{11} \times \det \mathbf{A}_{22\cdot 1} \tag{4.2.1a}$$

$$= \det \begin{bmatrix} \mathbf{I}_p & \mathbf{A}_{12} \\ \mathbf{0}_{(n-p),p} & \mathbf{A}_{22} \end{bmatrix} \times \det \begin{bmatrix} \mathbf{A}_{11\cdot 2} & \mathbf{0}_{p,(n-p)} \\ \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I}_{n-p} \end{bmatrix} = \det \mathbf{A}_{22} \times \det \mathbf{A}_{11\cdot 2} \tag{4.2.1b}$$

In (4.2.1a), $\mathbf{A}_{11}$ is assumed non-singular and so $\mathbf{A}_{11}^{-1}$ exists. Similarly, in (4.2.1b), $\mathbf{A}_{22}$ is assumed non-singular and so $\mathbf{A}_{22}^{-1}$ exists.

In the special case where $\mathbf{A}_{11} = \mathbf{I}_p$ and $\mathbf{A}_{22} = \mathbf{I}_{n-p}$, so that both matrices are non-singular with determinants of unity, combining (4.2.1a) and (4.2.1b) yields the important result

$$\det \mathbf{A}_{11\cdot 2} = \det \mathbf{A}_{22\cdot 1} \implies \det(\mathbf{I}_p - \mathbf{A}_{12}\mathbf{A}_{21}) = \det(\mathbf{I}_{n-p} - \mathbf{A}_{21}\mathbf{A}_{12}).$$

4.3 Partitioned Inverse

There are many different representations for the inverse of a partitioned matrix, some of which we will give here.[1] There are many ways to proceed. For example, we know that if $\mathbf{A}^{-1}$ exists then it must satisfy $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$. Partitioning $\mathbf{A}^{-1}$ so that it is conformable to post-multiply $\mathbf{A}$, and writing

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix},$$

then $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$ yields a set of 4 equations in the 4 unknowns:[2]

$$\mathbf{A}_{11}\mathbf{A}^{11} + \mathbf{A}_{12}\mathbf{A}^{21} = \mathbf{I}_p \tag{4.3.1a}$$
$$\mathbf{A}_{11}\mathbf{A}^{12} + \mathbf{A}_{12}\mathbf{A}^{22} = \mathbf{0}_{p\times(n-p)} \tag{4.3.1b}$$
$$\mathbf{A}_{21}\mathbf{A}^{11} + \mathbf{A}_{22}\mathbf{A}^{21} = \mathbf{0}_{(n-p)\times p} \tag{4.3.1c}$$
$$\mathbf{A}_{21}\mathbf{A}^{12} + \mathbf{A}_{22}\mathbf{A}^{22} = \mathbf{I}_{n-p} \tag{4.3.1d}$$

[1] A useful resource on inverting a partitioned matrix is Henderson and Searle (1981) which contains many of the results given here.

[2] Note that the answers obtained here are identical to those that we would obtain had we instead solved the equations arising from $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$.

How we solve the equations (4.3.1) depends upon what we want to assume. For example, if we assume that $\mathbf{A}_{11}$ is non-singular, without imposing a similar restriction on $\mathbf{A}_{22}$ then one solution is as follows: From (4.3.1a), we obtain

$$\mathbf{A}^{11} = \mathbf{A}_{11}^{-1}(\mathbf{I}_p - \mathbf{A}_{12}\mathbf{A}^{21}).$$

Substituting this result into (4.3.1c) yields, after some re-arrangement,

$$\mathbf{A}^{21} = -\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}$$

and hence

$$\mathbf{A}^{11} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}.$$

Similarly, from (4.3.1b), $\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22}$ which, on substitution into (4.3.1d) yields

$$\mathbf{A}^{22} = \mathbf{A}_{22\cdot 1}^{-1}$$

and hence

$$\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{21}\mathbf{A}_{22\cdot 1}^{-1}.$$

Gathering these results yields

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1} \\ -\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{A}_{22\cdot 1}^{-1} \end{bmatrix} \tag{4.3.2}$$

Alternatively, one might assume that $\mathbf{A}_{22}$ is non-singular, without imposing a similar restriction on $\mathbf{A}_{11}$, whereupon similar arguments to those above yield

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11\cdot 2}^{-1} & -\mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix} \tag{4.3.3}$$

If we assume that both $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are non-singular then both representations of the inverse are simultaneously valid and so comparison of (4.3.2) and (4.3.3) yields the following results:[3]

$$\mathbf{A}_{11\cdot 2}^{-1} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} \tag{4.3.4a}$$
$$\mathbf{A}_{22\cdot 1}^{-1} = \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}. \tag{4.3.4b}$$

and

$$\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1} = \mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} \tag{4.3.4c}$$
$$\mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1}. \tag{4.3.4d}$$

Exploiting these results yields a variety of different expressions for $\mathbf{A}^{-1}$. For example, using (4.3.4b) in (4.3.3) yields

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11\cdot 2}^{-1} & -\mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1} & \mathbf{A}_{22\cdot 1}^{-1} \end{bmatrix}. \tag{4.3.5}$$

If we apply (4.3.4d) to (4.3.5) then we obtain

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11\cdot 2}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1} & \mathbf{A}_{22\cdot 1}^{-1} \end{bmatrix},$$

[3] Alternatively, we can apply results from Section 4.4 to the definitions of (4.1.2).

whereas the use of (4.3.4b) and (4.3.4c) in (4.3.3) yields

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11\cdot 2}^{-1} & -\mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{A}_{22\cdot 1}^{-1} \end{bmatrix}. \tag{4.3.6}$$

Finally, if we assume that all of $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ are non-singular, which implies that they are all square and of the same dimension, then

$$\begin{aligned} \mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} &= -(\mathbf{A}_{21} - \mathbf{A}_{22}\mathbf{A}_{12}^{-1}\mathbf{A}_{11})^{-1} \\ \mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} &= -(\mathbf{A}_{12} - \mathbf{A}_{11}\mathbf{A}_{21}^{-1}\mathbf{A}_{22})^{-1} \end{aligned}$$

whereupon (4.3.6) becomes

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11\cdot 2}^{-1} & (\mathbf{A}_{21} - \mathbf{A}_{22}\mathbf{A}_{12}^{-1}\mathbf{A}_{11})^{-1} \\ (\mathbf{A}_{12} - \mathbf{A}_{11}\mathbf{A}_{21}^{-1}\mathbf{A}_{22})^{-1} & \mathbf{A}_{22\cdot 1}^{-1} \end{bmatrix}.$$

In summary, we have seen 5 different expressions for the inverse of a partitioned matrix, with the validity of each expression depending upon what assumptions you are prepared to make about the non-singularity of the various sub-matrices of $\mathbf{A}$. Sometimes your assumptions will imply that more than one of these expression is a valid representation of the inverse, in which case you should work with the one that is most convenient.

Note that alternate approach would have been based around the decompositions (4.1.1). Noting that $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ we see that

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{I}_p & \mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{0}_{n-p,p} & \mathbf{A}_{22\cdot 1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0}_{p,n-p} \\ \mathbf{A}_{21} & \mathbf{I}_{n-p} \end{bmatrix}^{-1} \tag{4.3.7a}$$

$$= \begin{bmatrix} \mathbf{A}_{11\cdot 2} & \mathbf{0}_{p,n-p} \\ \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I}_{n-p} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I}_p & \mathbf{A}_{12} \\ \mathbf{0}_{n-p,p} & \mathbf{A}_{22} \end{bmatrix}^{-1}. \tag{4.3.7b}$$

In these decompositions we see that we need inverses of the generic forms

$$\begin{bmatrix} \mathbf{A} & \mathbf{0}_{p,n-p} \\ \mathbf{B} & \mathbf{I}_{n-p} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mathbf{I}_p & \mathbf{A} \\ \mathbf{0}_{n-p,p} & \mathbf{B} \end{bmatrix}.$$

Without repeating the steps used in finding our initial expressions for the inverse we have, from (4.3.3) and (4.3.2) respectively,

$$\begin{bmatrix} \mathbf{A} & \mathbf{0}_{p,n-p} \\ \mathbf{B} & \mathbf{I}_{n-p} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0}_{p,n-p} \\ -\mathbf{B}\mathbf{A}^{-1} & \mathbf{I}_{n-p} \end{bmatrix} \tag{4.3.8}$$

and

$$\begin{bmatrix} \mathbf{I}_p & \mathbf{A} \\ \mathbf{0}_{n-p,p} & \mathbf{B} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}_p & -\mathbf{A}\mathbf{B}^{-1} \\ \mathbf{0}_{n-p,p} & \mathbf{B}^{-1} \end{bmatrix}. \tag{4.3.9}$$

Using these results in conjunction with (4.3.7), we see that

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{I}_p & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1} \\ \mathbf{0}_{n-p,p} & \mathbf{A}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0}_{p,n-p} \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I}_{n-p} \end{bmatrix} \tag{4.3.10a}$$

$$= \begin{bmatrix} \mathbf{A}_{11\cdot 2}^{-1} & \mathbf{0}_{p,n-p} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1} & \mathbf{I}_{n-p} \end{bmatrix} \begin{bmatrix} \mathbf{I}_p & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0}_{n-p,p} & \mathbf{A}_{22}^{-1} \end{bmatrix}. \tag{4.3.10b}$$

On completing the multiplications we see that (4.3.10a) yields (4.3.2) and (4.3.10b) yields (4.3.3).

4.4 Schur Complements

In Section 4.3 we saw repeated appearances of terms of the form $\mathbf{A}_{11\cdot 2}$ or $\mathbf{A}_{22\cdot 1}$. Such terms are known as Schur complements.[4] We have already seen one on the most important results for Schur complements in equations (4.3.4a) and (4.3.4b). Nevertheless, the literature on Schur complements is far more wide-ranging covering, inter alios, topics such as rank, inertia, and inequalities. Ouellette (1981) provides an exhaustive and accessible treatment of Schur complements and is highly recommended.

One might think of a Schur complement as a special case of the matrix sum $\mathbf{A} + \mathbf{BCD}$, where here $\mathbf{A}$ is assumed to be a non-singular matrix of dimension $n \times n$, and $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ of dimensions $n \times p$, $p \times q$, and $q \times n$, respectively. In particular, no assumption is made of either the squareness or non-singularity of $\mathbf{C}$ (or $\mathbf{B}$ or $\mathbf{D}$ for that matter). Within this framework, Henderson and Searle (1981, equations (21)–(26)) provide the following set of results for an inverse:

$$\begin{aligned}
(\mathbf{A}+\mathbf{BCD})^{-1} &= \mathbf{A}^{-1} - (\mathbf{I}_n + \mathbf{A}^{-1}\mathbf{BCD})^{-1}\mathbf{A}^{-1}\mathbf{BCDA}^{-1} && (4.4.1a)\\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{I}_n + \mathbf{BCDA}^{-1})^{-1}\mathbf{BCDA}^{-1} && (4.4.1b)\\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I}_p + \mathbf{CDA}^{-1}\mathbf{B})^{-1}\mathbf{CDA}^{-1} && (4.4.1c)\\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BC}(\mathbf{I}_q + \mathbf{DA}^{-1}\mathbf{BC})^{-1}\mathbf{DA}^{-1} && (4.4.1d)\\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BCD}(\mathbf{I}_n + \mathbf{A}^{-1}\mathbf{BCD})^{-1}\mathbf{A}^{-1} && (4.4.1e)\\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BCDA}^{-1}(\mathbf{I}_n + \mathbf{BCDA}^{-1})^{-1}. && (4.4.1f)
\end{aligned}$$

Clearly, these results could be combined with our earlier results on the inverse of a partitioned matrix to yield further expressions for the inverse of a partitioned matrix.

From either (4.4.1c) or (4.4.1d), if $\mathbf{C}$ is also non-singular we obtain the result

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}, \qquad (4.4.2)$$

which is known variously as the *matrix inversion lemma*, the *Woodbury matrix identity*, the *Sherman-Morrison-Woodbury formula*, or just the *Woodbury formula*.[5]

4.5 Some Properties of Triangular Matrices

Triangular matrices can be thought of as special cases of partitioned matrices. However, they occur sufficiently frequently that it is worth a few brief words to highlight some of their properties rather than having to simplify those obtained previously for partitioned matrices. For the most part we will deal with lower triangular matrices, with the analysis of upper triangular matrices being completely analogous. We shall start by demonstrating that the inverse of a lower triangular matrix is itself lower triangular. First, let the $n \times n$ matrix $\mathbf{L}_n$ be lower triangular

[4] Strictly speaking the Schur complements can be defined with respect to any non-singular sub-matrix in $\mathbf{A}$; see Ouellette (1981, equation (1.6)). However, for any such sub-matrix there will be a set of elementary matrix operations, notionally both row and column operations involving permutation matrices, that will make it possible to reduce the problem to that considered here.

[5] See https://en.wikipedia.org/wiki/Woodbury_matrix_identity.

and define the $(n+1)\times(n+1)$ lower triangular matrix

$$\mathbf{L}_{n+1} = \begin{bmatrix} \mathbf{L}_n & \mathbf{0}_{n,1} \\ \mathbf{x}' & \alpha \end{bmatrix} = \alpha \begin{bmatrix} \alpha^{-1}\mathbf{L}_n & \mathbf{0}_{n,1} \\ \alpha^{-1}\mathbf{x}' & 1 \end{bmatrix},$$

where the second equality assumes that $\alpha \neq 0$. From the first equality we see that if $\alpha = 0$ then the final column of $\mathbf{L}_{n+1}$ is comprised entirely of zeros. Expansion of $\det \mathbf{L}_{n+1}$ along this final column yields $\det \mathbf{L}_{n+1} = 0$, making $\mathbf{L}_{n+1}$ singular. Hence, the non-singularity of $\mathbf{L}_{n+1}$ requires $\alpha \neq 0$, which shall be assumed hereafter. Given this assumption, equation (4.2.1b) implies that $\det \mathbf{L}_{n+1} = \alpha \det \mathbf{L}_n$. As the non-singularity of $\mathbf{L}_{n+1}$ requires $\det \mathbf{L}_{n+1} \neq 0$, and as $\alpha \neq 0$ by assumption, it follows that $\det \mathbf{L}_n \neq 0$ which, in turn, implies that $\mathbf{L}_n$ must also be non-singular in order for $\mathbf{L}_{n+1}$ to be non-singular. The non-singularity of $\mathbf{L}_n$ will also be assumed hereafter.

Next, from equation (4.3.8), we have

$$\mathbf{L}_{n+1}^{-1} = \frac{1}{\alpha} \begin{bmatrix} \alpha\mathbf{L}_n^{-1} & \mathbf{0}_{n,1} \\ -\mathbf{x}'\mathbf{L}_n^{-1} & 1 \end{bmatrix},$$

and so it follows that if $\mathbf{L}_n^{-1}$ is lower triangular then so too is $\mathbf{L}_{n+1}^{-1}$.

Finally, consider the case of a 2×2 non-singular lower triangular matrix:[6]

$$\mathbf{L}_2 = \begin{bmatrix} a & 0 \\ b & c \end{bmatrix}.$$

Clearly, non-singularity of $\mathbf{L}_2$ requires $\det \mathbf{L}_2 = ac \neq 0$ which implies that $a \neq 0$ and $c \neq 0$. In Example 1.3.8 we derived the inverse of a (2×2) matrix. From this result we see that the inverse of $\mathbf{L}_2$ is

$$\mathbf{L}_2^{-1} = \frac{1}{ac} \begin{bmatrix} c & 0 \\ -b & a \end{bmatrix},$$

which is also lower triangular. We have already established that if $\mathbf{L}_n^{-1}$ is lower triangular then so too is $\mathbf{L}_{n+1}^{-1}$. Consequently, given that $\mathbf{L}_2^{-1}$ is lower triangular it follows that $\mathbf{L}_n^{-1}$ is lower triangular for $n = 2, 3, 4, \ldots$, as required.

4.6 More On Solving Systems of Linear Equations

Our most systematic treatment of solving linear equations to date was that of Section 2.5.2, which cast the analysis in terms of the relative ranks of the coefficient and augmented coefficient matrices of the system of equations. In this section we provide ageneral solution. We will summarize the results in a set of theorems.

Theorem 4.1 (Consistency)**.** *A set of linear equations can be solved, if and only if, they are consistent.* □

So what does consistency require?

[6] In the $n = 1$ case, a scalar is trivially lower triangular but it is also trivially pretty much everything else and so we are better off treating the $n = 2$ case as the base case.

Theorem 4.2 (Rank Condition for Consistency of Equations)**.** *The equations* $\mathbf{Ax} = \mathbf{y}$ *are consistent if, and only if, the rank of the augmented coefficient matrix* $[\mathbf{A}, \mathbf{y}]$ *is equal to the rank of* $\mathbf{A}$*.* □

Corollary 4.1 (An Alternative Statement of Theorem 4.2)**.** *If* $\mathbf{A}$ *is an* $m \times n$ *matrix of rank* $\rho(\mathbf{A}) \le m$*, then the equations* $\mathbf{Ax} = \mathbf{y}$ *are consistent if, and only if, the last* $m - \rho(\mathbf{A})$ *rows of* $\mathbf{P}[\mathbf{A}, \mathbf{y}]$ *are comprised entirely of zeros, where* $\mathbf{P}$ *is the matrix that reduces* $[\mathbf{A}, \mathbf{y}]$ *to row echelon form.* □

Theorem 4.3 (A General Solution)**.** *Let* $\mathbf{Ax} = \mathbf{y}$ *be a consistent set of equations where the* $m \times n$ *coefficient matrix has rank* $r = \rho(\mathbf{A}) \le \min(m, n)$*. Moreover, let the equations be ordered such that in the partitioning*

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

the $r \times r$ *matrix* $\mathbf{A}_{11}$ *is non-singular. Partition* $\mathbf{x} = [\mathbf{x}_1'\mathbf{x}_2]'$ *and* $\mathbf{y} = [\mathbf{y}_1'\mathbf{y}_2]'$ *conformably. Then a parametric solution to the equations is given by*

$$\mathbf{x}_1 = \mathbf{A}_{11}^{-1}(\mathbf{y}_1 - \mathbf{A}_{12}\mathbf{x}_2). \tag{4.6.1}$$

for $\mathbf{x}_2$ *an arbitrary* $(n - r)$*-vector.* □

A proof of this result is instructive.

Proof. Given our assumptions on $\mathbf{A}$, it must be the case that the rows $[\mathbf{A}_{21}, \mathbf{A}_{22}]$ can be written as linear combinations of $[\mathbf{A}_{11}, \mathbf{A}_{12}]$. That is, there exists some matrix $\mathbf{F}$, say, such that

$$[\mathbf{A}_{21}, \mathbf{A}_{22}] = \mathbf{F}[\mathbf{A}_{11}, \mathbf{A}_{12}].$$

Equally, there must exist some matrix $\mathbf{J}$, say, such that

$$\begin{bmatrix} \mathbf{A}_{12} \\ \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} \\ \mathbf{A}_{21} \end{bmatrix} \mathbf{J}.$$

Combining these two observations we see that (i) $\mathbf{A}_{21} = \mathbf{FA}_{11} \implies \mathbf{F} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$, (ii) $\mathbf{A}_{12} = \mathbf{A}_{11}\mathbf{J} \implies \mathbf{J} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, and (iii)

$$\mathbf{A}_{22} = \mathbf{FA}_{12} = \mathbf{FA}_{11}\mathbf{J} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{11}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}.$$

this allows us to write

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{11}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{11} & \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{11}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_r \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1} \end{bmatrix} \mathbf{A}_{11}[\mathbf{I}_r, \mathbf{A}_{11}^{-1}\mathbf{A}_{12}].$$

Now, if we partition both $\mathbf{x} = [\mathbf{x}_1', \mathbf{x}_2']'$ and $\mathbf{y} = [\mathbf{y}_1', \mathbf{y}_2']'$, so that both $\mathbf{x}_1$ and $\mathbf{y}_1$ contain r terms then

$$\mathbf{Ax} = \mathbf{y} \implies \begin{bmatrix} \mathbf{A}_{11}(\mathbf{x}_1 + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}_2) \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{11}(\mathbf{x}_1 + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{x}_2) \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}. \tag{4.6.2}$$

From the first row of (4.6.2) we infer that

$$\mathbf{x}_1 = \mathbf{A}_{11}^{-1}(\mathbf{y}_1 - \mathbf{A}_{12}\mathbf{x}_2), \tag{4.6.3}$$

as required. From the second row of (4.6.2) we see that the consistency of the equations implies that $\mathbf{y}_2 = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{y}_1$ must hold too. □

There is much that can be learned on inspection of (4.6.1). First, if $\mathbf{A}$ is non-singular then $\mathbf{A}_{11} \equiv \mathbf{A}$ and the terms $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ simply do not exist and neither do $\mathbf{x}_2$ nor $\mathbf{y}_2$. Hence, $\mathbf{y}_1 \equiv \mathbf{y}$ and $\mathbf{x}_1 \equiv \mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. That is, the solution is unique. Second, a unique solution will exist whenever the term $\mathbf{A}_{12}\mathbf{x}_2$ does not exist, which is a consequence of the coefficient matrix having full column rank. That is, if $r = n$ then there will be a unique solution to $\mathbf{Ax} = \mathbf{y}$, provided the system of equations is consistent. If, however, the coefficient matrix has less than full column rank then there will be an infinite number of solutions given by (4.6.1), where now $\mathbf{x}_2$ is arbitrary.

Example 4.6.1. A General Solution.
Consider the set of equations

$$2x_1 + 3x_2 + x_3 = 14 \tag{4.6.4a}$$
$$x_1 + x_2 + x_3 = 6 \tag{4.6.4b}$$
$$3x_1 + 5x_2 + x_3 = 22. \tag{4.6.4c}$$

It is straight-forward to reduce the augmented coefficient matrix for this system from

$$\begin{bmatrix} 2 & 3 & 1 & 14 \\ 1 & 1 & 1 & 6 \\ 3 & 5 & 1 & 22 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 2 & 4 \\ 0 & 1 & -1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

which implies the set of solutions $\widetilde{x}_1 = 4 - 2x_3$, $\widetilde{x}_2 = 2 + x_3$, with x_3 unconstrained: $\widetilde{x}_3 = x_3$.[7] Alternatively, we might have reduced the coefficient matrix $\mathbf{A}$ to row echelon form in order to determine rank:

$$\mathbf{PA} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -2 & 0 \\ -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 1 \\ 3 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix},$$

where $\mathbf{P}$ is the set of elementary row operations required to reduce $\mathbf{A}$ to row echelon form. We see that the rank of $\mathbf{A}$ is 2. If we then define $\mathbf{x}_1 = [x_1, x_2]'$, $\mathbf{x}_2 = x_3$, $\mathbf{y}_1 = [14, 6]'$, and partition $\mathbf{A}$ according to

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & a_{22} \end{bmatrix} = \left[\begin{array}{cc:c} 2 & 3 & 1 \\ 1 & 1 & 1 \\ \hdashline 3 & 5 & 1 \end{array}\right],$$

then, from (4.6.1), we have

$$\mathbf{x}_1 = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} 14 \\ 6 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} x_3 \right) = \frac{1}{(2)(1) - (3)(1)} \begin{bmatrix} 1 & -3 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 14 - x_3 \\ 6 - x_3 \end{bmatrix} = \begin{bmatrix} 4 - 2x_3 \\ 2 + x_3 \end{bmatrix},$$

[7] You will sometimes see such a parametric solution written

$$\begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \\ \widetilde{x}_3 \end{bmatrix} = \begin{bmatrix} 4 - t \\ 2 + t \\ t \end{bmatrix},$$

to highlight the fact that the solution involves an unspecified parameter t. Once a value for t is given then a specific solution of $\widetilde{\mathbf{x}} = (\widetilde{x}_1, \widetilde{x}_2, \widetilde{x}_3)$ is available.

which implies the solution vector, for any given x_3,

$$\tilde{\mathbf{x}} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} 4 - 2x_3 \\ 2 + x_3 \\ x_3 \end{bmatrix},$$

as before. □

It is worth noting that there is nothing special about the way in which the equations $\mathbf{Ax} = \mathbf{y}$ need be ordered beyond the requirement that we have a non-singular top left block of dimension $\rho(\mathbf{A}) \times \rho(\mathbf{A})$. In the previous example, there are 6 (3!) different ways in which we could have ordered the equations and the result would have been unaffected.

Example 4.6.2. Example 4.6.1 Re-visited and Re-ordered.
Suppose, instead, that the equations of the previous example had been ordered as follows:

$$\begin{aligned} 3x_1 + 5x_2 + x_3 &= 22 \\ x_1 + x_2 + x_3 &= 6 \\ 2x_1 + 3x_2 + x_3 &= 14. \end{aligned}$$

The rank of any matrix is unaffected by the interchange of rows and so the rank of the coefficient matric in this system of equations is also 2. Writing

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} 3 & 5 & \vdots & 1 \\ 1 & 1 & \vdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 2 & 3 & \vdots & 1 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_3 \end{bmatrix} = \begin{bmatrix} 22 \\ 6 \\ \cdots \\ 14 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_3 \end{bmatrix} = \begin{bmatrix} 22 \\ 6 \\ \cdots \\ 14 \end{bmatrix},$$

we have

$$\begin{aligned} \mathbf{x}_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 3 & 5 \\ 1 & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} 22 \\ 6 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} x_3 \right) = \frac{1}{(3)(1) - (5)(1)} \begin{bmatrix} 1 & -5 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 22 - x_3 \\ 6 - x_3 \end{bmatrix} \\ &= \begin{bmatrix} 4 - 2x_3 \\ 2 + x_3 \end{bmatrix}, \end{aligned}$$

as before, and which agin implies the solution vector, for any given x_3, is

$$\tilde{\mathbf{x}} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} 4 - 2x_3 \\ 2 + x_3 \\ x_3 \end{bmatrix},$$

as before. □

4.7 The Frisch-Waugh-Lovell Theorem

One of our most common applications of partitioned matrices arises in the context of the linear regression model, where we wish to partition the explanatory variables into two distinct groups. Specifically, we have in mind taking a linear regression model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{4.7.1}$$

where it is assumed that $\mathbf{X}$ has full column rank, partitioning the matrix of explanatory variables into two group $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ and partitioning the vector of coefficients conformably so that $\boldsymbol{\beta} = [\boldsymbol{\beta}_1', \boldsymbol{\beta}_2']'$ so that

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}. \tag{4.7.2}$$

Note that equations (4.7.1) and (4.7.2) are identical, it is only the way that we have chosen to write the equations that differs between the two. The least squares estimator in either case is given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(y - \mathbf{X}\boldsymbol{\beta})'(y - \mathbf{X}\boldsymbol{\beta}) \tag{4.7.3}$$

and

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \underset{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}{\operatorname{argmin}}(y - \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2)'(y - \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2). \tag{4.7.4}$$

From (4.7.3), we see that the relevant first-order conditions are given by

$$\mathbf{0} = \frac{\boldsymbol{\partial}}{\boldsymbol{\partial\beta}}[\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \implies \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

From (4.7.3), we have

$$\begin{aligned}
\mathbf{0} &= \begin{bmatrix} \frac{\boldsymbol{\partial}}{\boldsymbol{\partial\beta}_1}[\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}_1'\mathbf{X}_1'\mathbf{y} - 2\boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{y} - 2\boldsymbol{\beta}_1'\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\beta}_1'\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{X}_2\boldsymbol{\beta}_2] \\ \frac{\boldsymbol{\partial}}{\boldsymbol{\partial\beta}_2}[\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}_1'\mathbf{X}_1'\mathbf{y} - 2\boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{y} - 2\boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\beta}_1'\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{X}_2\boldsymbol{\beta}_2] \end{bmatrix}_{\substack{\boldsymbol{\beta}_1=\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{\beta}_2=\hat{\boldsymbol{\beta}}_2}} \\
&= \begin{bmatrix} -2\mathbf{X}_1'\mathbf{y} - 2\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + 2\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 \\ -2\mathbf{X}_2'\mathbf{y} - 2\mathbf{X}_2'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + 2\mathbf{X}_2'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 \end{bmatrix}
\end{aligned}$$

or simply

$$\begin{aligned}
\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix} \\
\implies \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},
\end{aligned}$$

and so the two approaches are, as one might expect, giving you exactly the same least squares estimators.

Combining the definitions of (4.1.2) with (4.3.6), we see that

$$\begin{aligned}
&\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \\
&\quad = \begin{bmatrix} (\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1)^{-1} & -(\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1} \\ -(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1} & (\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1} \end{bmatrix},
\end{aligned}$$

with $\mathbf{M}_{X_j} = \mathbf{I}_n - \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$, $j \in \{1, 2\}$, because

$$\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1 = \mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1$$

and

$$\mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2.$$

Thus,

$$\begin{aligned}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} &= \begin{bmatrix}(\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2')\mathbf{y}\\ (\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'(\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}\end{bmatrix}\\ &= \begin{bmatrix}(\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{y}\\ (\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{y}\end{bmatrix}.\end{aligned} \tag{4.7.5}$$

From the idempotency of the matrices $\mathbf{M}_{X_j}$, $j \in \{1, 2\}$, we can write

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \begin{bmatrix}(\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{M}_{X_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{M}_{X_2}\mathbf{y}\\ (\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{M}_{X_1}\mathbf{y}\end{bmatrix}$$

from which we see that identical values for $\hat{\boldsymbol{\beta}}_1$ are obtained from the two regression models (4.7.1) and

$$\mathbf{M}_{X_2}\mathbf{y} = \mathbf{M}_{X_2}\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{M}_{X_2}\mathbf{u} \tag{4.7.6}$$

Observe that $\mathbf{M}_{X_2}\mathbf{y}$ are the least squares residuals from the regression of $\mathbf{y}$ on $\mathbf{X_2}$ and that $\mathbf{M}_{X_2}\mathbf{X}_1$ are the least squares residuals from the multivariate regression of $\mathbf{X}_1$ on $\mathbf{X_2}$. Also note that the pre-multiplication of the equation by $\mathbf{M}_{X_2}$ results in the disturbance term being multiplied by the same quantity but, as this is an unobserveable, it is of no relevance to the least squares estimation. Similarly, identical values for $\hat{\boldsymbol{\beta}}_2$ are obtained from the two regression models (4.7.1) and

$$\mathbf{M}_{X_1}\mathbf{y} = \mathbf{M}_{X_1}\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_{X_1}\mathbf{u} \tag{4.7.7}$$

Observe that $\mathbf{M}_{X_2}\mathbf{y}$ are the least squares residuals from the regression of $\mathbf{y}$ on $\mathbf{X_1}$ and that $\mathbf{M}_{X_2}\mathbf{X}_1$ are the least squares residuals from the multivariate regression of $\mathbf{X}_2$ on $\mathbf{X_1}$. Perhaps most remarkably the least squares residuals from (4.7.1), (4.7.6), and (4.7.7) are also identical. To see this last result observe that the residuals from (4.7.1) are given by the expression

$$\mathbf{e} = \mathbf{M}_X\mathbf{y}. \tag{4.7.8}$$

If we define $\mathbf{W}_1 = \mathbf{M}_{X_2}\mathbf{X}_1$ and $\mathbf{W}_2 = \mathbf{M}_{X_1}\mathbf{X}_2$, respectively, then we see that the residuals from equation (4.7.6) are given by

$$\mathbf{e}_1 = \mathbf{M}_{W_1}\mathbf{M}_{X_2}\mathbf{y}.$$

Similarly, the residuals from equation (4.7.7) are given by

$$\mathbf{e}_2 = \mathbf{M}_{W_2}\mathbf{M}_{X_1}\mathbf{y}.$$

Clearly, we need to show that $\mathbf{M}_X = \mathbf{M}_{W_1}\mathbf{M}_{X_2} = \mathbf{M}_{W_2}\mathbf{M}_{X_1}$. The easiest way to do this is to start with $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We might choose to start with start with (4.3.2), in which case

$$\mathbf{P}_X = \mathbf{P}_{X_1} \underbrace{+\mathbf{P}_{X_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{P}_{X_1} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{P}_{X_1}}_{=-\mathbf{M}_{X_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{P}_{X_1}}$$

$$\underbrace{-\mathbf{P}_{X_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2' + \mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'}_{=+\mathbf{M}_{X_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'}$$
$$= \mathbf{P}_{X_1} + \mathbf{M}_{X_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_{X_1}. \tag{4.7.9}$$

Alternatively one might start with (4.3.3), whereupon we see that

$$\mathbf{P}_X = \mathbf{P}_{X_2} + \mathbf{M}_{X_2}\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_{X_2}. \tag{4.7.10}$$

Note that this result can be obtained purely by symmetry arguments from (4.7.9) because it simply amounts to interchanging the columns $\mathbf{X}_1$ and $\mathbf{X}_2$ in $\mathbf{X}$. Note that in terms of the discussion of Section 1.5, we can think of a projection on to the space spanned by the columns of $\mathbf{X}$ as being comprised of two parts. First, there is a projection onto the the space spanned by a subset of the columns of $\mathbf{X}$, be that $\mathbf{X}_1$ or $\mathbf{X}_2$, and added to that is a projection onto that part of the remaining columns of $\mathbf{X}$ that is orthogonal to the initial projection being either the column space of $\mathbf{M}_{X_1}\mathbf{X}_2$ or that of $\mathbf{M}_{X_2}\mathbf{X}_1$, respectively. Now that we have these results it is a simple matter to find the projections onto the orthogonal complements, or the null spaces of subsets of columns of $\mathbf{X}$. That is,

$$\begin{aligned}\mathbf{M}_X = \mathbf{I}_n - \mathbf{P}_X &= \mathbf{M}_{X_1} - \mathbf{M}_{X_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_{X_1}\\ &= \mathbf{M}_{X_2} - \mathbf{M}_{X_2}\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_{X_2}.\end{aligned}$$

These results are so helpful in so many contexts that it is worthwhile gathering them together in a theorem.

Theorem 4.4 (Projections Onto Subsets of Columns of a Matrix)**.** *Let* $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ *be a matrix of full column rank and, for any* $n \times k$ *matrix* $\mathbf{A}$ *of full column rank, define* $\mathbf{P}_A = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}$ *and* $\mathbf{M}_A = \mathbf{I}_n - \mathbf{P}_A$*. Then*

$$\mathbf{P}_{X_1} = \mathbf{P}_X - \mathbf{M}_{X_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_{X_1} \tag{4.7.11a}$$
$$\mathbf{P}_{X_2} = \mathbf{P}_X - \mathbf{M}_{X_2}\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_{X_2} \tag{4.7.11b}$$
$$\mathbf{M}_{X_1} = \mathbf{M}_X + \mathbf{M}_{X_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{X_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_{X_1} \tag{4.7.11c}$$
$$\mathbf{M}_{X_2} = \mathbf{M}_X + \mathbf{M}_{X_2}\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_{X_2} \tag{4.7.11d}$$

are the various projections onto the spaces spanned by subsets of the columns of $\mathbf{X}$ *and also onto those subsets of the null sets of these columns that are spanned by the remaining columns of* $\mathbf{X}$. □

Armed with these results we are now in a position to establish the remainder of the Frisch-Waugh-Lovel Theorem. We have see that there are multiple ways in which we can generate the least squares estimates of a subset of the coefficients in a linear regression model. The remaining part of the theorem involves showing that the residuals obtained from these equations are all the same. As noted previously, we can do this by showing that $\mathbf{M}_X = \mathbf{M}_{W_1}\mathbf{M}_{X_2} = \mathbf{M}_{W_2}\mathbf{M}_{X_1}$, where $\mathbf{W}_1 = \mathbf{M}_{X_2}\mathbf{X}_1$ and $\mathbf{W}_2 = \mathbf{M}_{X_1}\mathbf{X}_2$. From Theorem 4.4 we have immediately the results we need. For example, from (4.7.11b), we see that $\mathbf{P}_{W_1} = \mathbf{P}_X - \mathbf{P}_{X_2}$, so that $\mathbf{M}_{W_1} = \mathbf{M}_X + \mathbf{P}_{X_2}$. Therefore,

$$\mathbf{M}_{W_1}\mathbf{M}_{X_2} = (\mathbf{M}_X + \mathbf{P}_{X_2})\mathbf{M}_{X_2} = \mathbf{M}_X\mathbf{M}_{X_2} = \mathbf{M}_X - \mathbf{M}_X\mathbf{P}_{X_2} = \mathbf{M}_X,$$

as required. Similarly, from (4.7.11a), $\mathbf{P}_{W_2} = \mathbf{P}_X - \mathbf{P}_{X_1}$ and so $\mathbf{M}_{W_2} = \mathbf{M}_X + \mathbf{P}_{X_1}$, which yields

$$\mathbf{M}_{W_2}\mathbf{M}_{X_1} = (\mathbf{M}_X + \mathbf{P}_{X_1})\mathbf{M}_{X_1} = \mathbf{M}_X\mathbf{M}_{X_1} = \mathbf{M}_X - \mathbf{M}_X\mathbf{P}_{X_1} = \mathbf{M}_X.$$

All these results can be combined into a version of the Frisch-Waugh-Lovell Theorem that is essentially that of Davidson and MacKinnon (2004, Theorem 2.1); some history of the result is available there.

Theorem 4.5 (Frisch-Waugh-Lovell Theorem)**.** *Consider the three equations*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \tag{4.7.12a}$$
$$\mathbf{M}_{X_2}\mathbf{y} = \mathbf{M}_{X_2}\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{M}_{X_2}\mathbf{u} \tag{4.7.12b}$$
$$\mathbf{M}_{X_1}\mathbf{y} = \mathbf{M}_{X_1}\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_{X_1}\mathbf{u} \tag{4.7.12c}$$

where the dependent variable $\mathbf{y}$ *and the unobserved disturbance* $\mathbf{u}$ *are* n*-vectors,* $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ *is an* $(n \times k)$ *matrix of explanators of full column rank, and* $\boldsymbol{\beta} = [\boldsymbol{\beta}_1', \boldsymbol{\beta}_2']'$ *is a* k*-vector of parameters that has been partitioned conformably with* $\mathbf{X}$*. For any matrix* $(n \times k)$ *matrix* $\mathbf{A}$ *of full column rank, define* $\mathbf{M}_A = \mathbf{I}_n - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$*, where* $\mathbf{I}_n$ *denotes an identity matrix of order* n*. Then*

(i) The least squares estimates of $\boldsymbol{\beta}_1$ *obtained from equations* (4.7.12a) *and* (4.7.12b) *are numerically identical, as are the least squares estimates of* $\boldsymbol{\beta}_2$ *obtained from equations* (4.7.12a) *and* (4.7.12c)*.*

(ii) The least squares residuals obtained from equations (4.7.12a)*,* (4.7.12b)*, and* (4.7.12c) *are numerically identical.* □

Two final observations. First, the Frisch-Waugh-Lovell Theorem is a useful device that has application in both the theory of estimation and in the construction of test statistics, indeed anywhere that one may be interested in dealing with just a subset of the regression coefficients in a linear model. Second, a variant on Theorem 4.5(i) is to leave the dependent variable unchanged. That is, the estimation results would apply equally to the equations (4.7.12a) and (4.7.13):

$$\mathbf{y} = \mathbf{M}_{X_2}\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{M}_{X_2}\mathbf{u} \tag{4.7.13a}$$

and

$$\mathbf{y} = \mathbf{M}_{X_1}\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_{X_1}\mathbf{u}. \tag{4.7.13b}$$

However, this observation doesn't carry over to the residuals. Again using our definitions of $\mathbf{W}_1 = \mathbf{M}_{X_2}\mathbf{X}_1$ and $\mathbf{W}_2 = \mathbf{M}_{X_1}\mathbf{X}_2$, we see that the least squares residual from (4.7.12b) can be written

$$\mathbf{M}_{W_1}\mathbf{M}_{X_2}\mathbf{y} = \mathbf{M}_{X_2}\mathbf{y} - \mathbf{P}_{W_1}\mathbf{M}_{X_2}\mathbf{y} = \mathbf{M}_{X_2}\mathbf{M}_{W_1}\mathbf{y},$$

as $\mathbf{W}_1'\mathbf{M}_{X_2} = \mathbf{X}_1'\mathbf{M}_{X_2}\mathbf{M}_{X_2} = \mathbf{X}_1'\mathbf{M}_{X_2} = \mathbf{W}_1'$. Conversely, the residuals from (4.7.13a) are simply $\mathbf{M}_{W_1}\mathbf{y}$, which clearly differs from $\mathbf{M}_{X_2}\mathbf{M}_{W_1}\mathbf{y}$. Similarly, $\mathbf{M}_{W_2}\mathbf{y} \neq \mathbf{M}_{X_1}\mathbf{M}_{W_2}\mathbf{y}$ and so we conclude that neither of the equations (4.7.13) yield the same residuals as those of (4.7.12).

Chapter 5

The Generalized Inverse

5.1 The Generalized Inverse In General

In Section 1.3.7 we introduced the notion of the inverse of a non-singular matrix. That is, attention was restricted to the case where the matrix to be inverted was both square and of full rank. Obvious ways in which one may wish to relax these assumptions includes the consideration of square matrices with less than full rank and non-square matrices which may or may not have either full row rank or full column rank. In all cases we will assume that the matrix to be inverted has non-zero rank, i.e. it is a non-zero matrix.

Recall that if $\mathbf{A}$ is a square matrix of dimension $m \times m$ with rank m, so that it is non-singular, then there exists a unique matrix $\mathbf{A}^{-1}$, called the regular inverse of $\mathbf{A}$, or simply the inverse of $\mathbf{A}$, with the properties:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_m,$$

where $\mathbf{I}_m$ is an identity matrix of order m. We see that $\mathbf{A}^{-1}$ is both the right inverse and the left inverse of $\mathbf{A}$ because both post-multiplication of $\mathbf{A}$ by $\mathbf{A}^{-1}$ and pre-multiplication of $\mathbf{A}$ by $\mathbf{A}^{-1}$ yields the identity matrix. It is these properties that will form the basis of a definition of a generalized inverse, or g-inverse, of $\mathbf{A}$. In particular, either pre-multiplication by $\mathbf{A}$ of both sides of the equation $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_m$ or post-multiplication of $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_m$ by $\mathbf{A}$ yields the result

$$\mathbf{A}\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}.$$

This is the property that we use to define the generalized inverse of $\mathbf{A}$. Specifically, suppose that $\mathbf{A}$ is an $m \times n$ matrix of rank $r \leq \min(m, n)$. Then the matrix $\mathbf{A}^-$ will be a g-inverse of $\mathbf{A}$ if and only if it is an $n \times m$ matrix such that

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}. \tag{5.1.1}$$

That is, $\mathbf{A}$ may be either a square or a rectangular (not square) array. If square then the g-inverse is relevant if $\mathbf{A}$ is rank deficient. If rectangular then the g-inverse is relevant whether $\mathbf{A}$ has full row or column rank, or whether its rank is smaller than the smallest of its dimensions.

Among the implications of our definition is that $\mathbf{A}^-\mathbf{A}$ is idempotent (pre-multiply (5.1.1) by $\mathbf{A}^-$) and $\rho(\mathbf{A}^-\mathbf{A}) = \rho(\mathbf{A})$. Equally, $\mathbf{A}\mathbf{A}^-$ is idempotent (post-multiply (5.1.1) by $\mathbf{A}^-$) and $\rho(\mathbf{A}\mathbf{A}^-) = \rho(\mathbf{A})$, where $\rho(\mathbf{A})$ denotes the rank of $\mathbf{A}$.

There are many different ways of constructing a g-inverse. Consider for example the developments of Chapter 2. There we found that we could obtain the reduced row echelon form of a matrix, $\mathbf{A}$ say, by applying a series of elementary row operations only, which we may gather together in some matrix $\mathbf{P}$, say. It was often the case that if $\mathbf{A}$ had reduced rank that the reduced row echelon form had the following structure:

$$\mathbf{PA} = \begin{bmatrix} \mathbf{T} & \mathbf{F} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{5.1.2}$$

$$\mathbf{PA} = \begin{bmatrix} \mathbf{T} & \mathbf{F} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{T}$ is an upper triangular matrix with ones on the leading diagonal. In this case it can be shown that one possible g-inverse of $\mathbf{A}$ satisfying 5.1.1 is

$$\mathbf{G} = \begin{bmatrix} \mathbf{T}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{P}.$$

Note that we need enough rows of zeros at the bottom of the first matrix on the right-hand side of this equation to ensure that our g-inverse has the right dimensions, being the reverse of the dimensions of $\mathbf{A}$. Similarly we need enough zeros to the right of $\mathbf{T}^{-1}$ to ensure conformability under matrix multiplication with $\mathbf{P}$.

If elementary column operations are required to achieve the form on the right-hand side of 5.1.2, then these can be gathered together in a matrix $\mathbf{Q}$, say. In such cases we have

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{T} & \mathbf{F} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and a possible g-inverse is then

$$\mathbf{G} = \mathbf{Q} \begin{bmatrix} \mathbf{T}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{P}. \tag{5.1.3}$$

Example 5.1.1. Example 2.2.1 revisited.
Recall in Example 2.2.1, we reduced the matrix $\mathbf{A}$ to reduced row echelon form through elementary row operations gathered together in the matrix $\mathbf{P}$ (see Example 2.2.5). Specifically, if

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 11 \\ 3 & 7 & 22 \end{bmatrix} \qquad \text{and} \qquad \mathbf{P} = \frac{1}{11} \begin{bmatrix} 7 & -1 \\ -3 & 2 \end{bmatrix},$$

then we found that $\mathbf{PA} = [\mathbf{T}, \mathbf{F}]$, where $\mathbf{T} = \mathbf{I}_2$ and $\mathbf{F} = [5, 1]'$. Thus, a g-inverse of $\mathbf{A}$ is given by

$$\mathbf{G} = \begin{bmatrix} \mathbf{T}^{-1} \\ \mathbf{0} \end{bmatrix} \mathbf{P} = \begin{bmatrix} \mathbf{P} \\ 0\ 0 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 7 & -3 \\ -1 & 2 \\ 0 & 0 \end{bmatrix}.$$

We can check this:

$$\mathbf{AGA} = \frac{1}{11} \begin{bmatrix} 2 & 1 & 11 \\ 3 & 7 & 22 \end{bmatrix} \begin{bmatrix} 7 & -1 \\ -3 & 2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 & 11 \\ 3 & 7 & 22 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 11 \\ 3 & 7 & 22 \end{bmatrix}.$$

as required. □

Another form of g-inverse that is available when we can partition the $m \times n$ matrix $\mathbf{A}$ according to

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

with $\mathbf{A}_{11}$ non-singular and of dimension $\rho(\mathbf{A}) \times \rho(\mathbf{A})$, is

$$\mathbf{G} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{5.1.4}$$

where, if $\mathbf{A}$ is of dimension $m \times n$ then $\mathbf{G}$ contains enough zeros to have dimension $n \times m$.

Example 5.1.2. An Easy Generalized Inverse.
Suppose that

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 & 3 \\ 3 & -1 & 2 & -2 \\ 5 & 3 & 10 & 4 \end{bmatrix}.$$

It is straight-forward to show that $\mathbf{A}$ has rank $\rho(\mathbf{A}) = 2$,[1] and that

$$\mathbf{A}_{11}^{-1} = \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix}^{-1} = -\frac{1}{7}\begin{bmatrix} -1 & -2 \\ -3 & 1 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix}$$

Hence, a g-inverse of $\mathbf{A}$ is

$$\mathbf{G} = \frac{1}{7}\begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

You should check that this is so by substituting these definitions of $\mathbf{A}$ and $\mathbf{G}$ into the left-hand side of 5.1.1. □

For any matrix $\mathbf{A}$ of rank $\rho(\mathbf{A})$, whe can always use row and column operations to generate a non-singular $\rho(\mathbf{A}) \times \rho(\mathbf{A})$ matrix in the top left block of the matrix which suggests that some combination of (5.1.3) and (5.1.4) will allow us to find a g-inverse satisfying 5.1.1. Moreover, the previous two examples have illustrated that there are, in general, a variety of g-inverses available available.[2] To reinforce this point we have the following remarkable result.

Theorem 5.1 (The Family of Generalized Inverses)**.** *Let* $\mathbf{G}$ *denote any g-inverse of a matrix* $\mathbf{A}$*. Then the matrix* $\mathbf{G}^*$ *is also a g-inverse of* $\mathbf{A}$*, where*

$$\mathbf{G}^* = \mathbf{GAG} + (\mathbf{I} - \mathbf{GA})\mathbf{T} + \mathbf{S}(\mathbf{I} - \mathbf{AG}), \tag{5.1.5}$$

for any matrices $\mathbf{T}$ *and* $\mathbf{S}$ *of appropriate dimensions.* □

To see this, pre- and post-multiply $\mathbf{G}^*$ by $\mathbf{A}$.

What makes this result remarkable is the following observation. Suppose that $\mathbf{A}$ has some other generalized inverse $\widetilde{\mathbf{G}}$ say, that is different to $\mathbf{G}$. Setting $\mathbf{T} = \widetilde{\mathbf{G}}$ and $\mathbf{S} = \mathbf{GA}\widetilde{\mathbf{G}}$, we see that $\mathbf{G}^* = \widetilde{\mathbf{G}}$. That is, any other g-inverse of $\mathbf{A}$ can be generated from (5.1.5), which means that the entire set of g-inverses of $\mathbf{A}$ can be generated from (5.1.5), given appropriate choice of $\mathbf{T}$ and $\mathbf{S}$.

[1] Observe that the third row is equal to the second row plus twice the first row.

[2] Remember that if $\mathbf{A}$ is non-singular then there is only one possible g-inverse, the so-called regular inverse, which is always unique.

5.2 The Moore-Penrose Generalized Inverse

The g-inverse defined above is typically not unique unless, of course, $\mathbf{A}$ is non-singular in which case its inverse is uniquely defined. Indeed, if $\mathbf{A}$ has less than full rank then there is an infinite number of possible g-inverses. The most famous is the Moore-Penrose inverse, denoted $\mathbf{A}^+$, which adds some additional restrictions that make it unique. In particular, it is defined to satisfy the following 4 properties, which are known as the Penrose conditions:

(i) It is a generalized inverse, so that $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$;

(ii) It is reflexive, so that $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$;

(iii) $\mathbf{A}\mathbf{A}^+$ is symmetric; and

(iv) $\mathbf{A}^+\mathbf{A}$ is symmetric.

The Moore-Penrose inverse is attractive because it is relatively easy to construct. Every $m \times n$ matrix $\mathbf{A}$, of rank r admits a full rank factorization of the form $\mathbf{A} = \mathbf{KL}$ where $\mathbf{K}$ is $m \times r$ and $\mathbf{L}$ is $r \times n$. The Moore-Penrose inverse can then be constructed according to $\mathbf{A}^+ = \mathbf{L}'(\mathbf{K}'\mathbf{A}\mathbf{L}')^{-1}\mathbf{K}'$. We will illustrate this shortly (Example 7.2.2) but first we need to see how to construct a full rank factorization, which is considered in the next chapter. An alternative expression for the Moore-Penrose inverse is

$$\mathbf{A}^+ = \mathbf{A}'(\mathbf{A}\mathbf{A}')^-\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}',$$

where $\mathbf{B}^-$ denotes an arbitrary g-inverse of $\mathbf{B}$.

5.3 The Left and Right Inverse

If $\mathbf{A}$ is non-singular then it is square, has full rank, and it has a unique inverse, called the *regular inverse*, denoted $\mathbf{A}^{-1}$, such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. We see that $\mathbf{A}^{-1}$ acts as both a left inverse ($\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$) and a *right inverse* ($\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$). With rectangular matrices or square matrices with less than full rank, then it need not be the case that the left and right inverses are equal. We summarize the the key results in the following two theorems.

Theorem 5.2 (Left Inverse)**.** *Let $\mathbf{A}$ denote an $m \times n$ matrix. Then, a necessary and sufficient condition for the existence of an $n \times m$ matrix $\mathbf{L}$, called the left inverse of $\mathbf{A}$, such that $\mathbf{L}\mathbf{A} = \mathbf{I}_n$ is that $\mathbf{A}$ has full column rank. The corresponding order condition, which is necessary but not sufficient, is that $m \geq n$.* □

Theorem 5.3 (Right Inverse)**.** *Let $\mathbf{A}$ denote an $m \times n$ matrix. Then, a necessary and sufficient condition for the existence of an $n \times m$ matrix $\mathbf{R}$, called the right inverse of $\mathbf{A}$, such that $\mathbf{A}\mathbf{R} = \mathbf{I}_m$ is that $\mathbf{A}$ has full row rank. The corresponding order condition, which is necessary but not sufficient, is that $m \leq n$.* □

Theorem 5.4. *A matrix $\mathbf{A}$ has both a left inverse $\mathbf{L}$ and a right inverse $\mathbf{R}$ if, and only if, it is non-singular, in which case $\mathbf{L} = \mathbf{R} = \mathbf{A}^{-1}$ is the regular inverse of $\mathbf{A}$.*

Examples of left and right inverses that are commonly encountered are the following:

(i) Suppose $\mathbf{X}$ is $m \times n$ and of full column rank, so that $m \geq n$, then a left inverse of $\mathbf{X}$ is $\mathbf{L}_X = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(ii) Suppose $\mathbf{X}$ is $m \times n$ and of full row rank, so that $m \leq n$, then a right inverse of $\mathbf{X}$ is $\mathbf{R}_X = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}$.

Clearly the rank conditions on $\mathbf{X}$ are essential so that the regular inverses, $(\mathbf{X}'\mathbf{X})^{-1}$ and $(\mathbf{X}\mathbf{X}')^{-1}$, respectively, exist. It is straight-forward to show that, when defined, $\mathbf{L}_X$ and $\mathbf{R}_X$ are g-inverses.

5.4 Solving Linear Systems of Equations Using a Generalized Inverse

Our most systematic treatment of solving linear equations to date was that of Section 4.6. Having introduced the notion of a g-inverse it seems not unreasonable to ask whether or not a system of equations $\mathbf{A}\mathbf{x} = \mathbf{y}$ might have a solution of the form $\mathbf{x} = \mathbf{G}\mathbf{y}$, where $\mathbf{G}$ is a g-inverse of $\mathbf{A}$, by direct analogy to the solution of $\mathbf{A}\mathbf{x} = \mathbf{y}$ when $\mathbf{A}$ is non-singular, i.e., $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. Not surprisingly, perhaps, the answer is yes. We provide the details below.

Theorem 5.5 (Expressing the Solution in Terms of a Generalized Inverse)**.** *The consistent equations* $\mathbf{A}\mathbf{x} = \mathbf{y}$ *for* $\mathbf{y} \neq \mathbf{0}$ *have a solution* $\mathbf{x} = \mathbf{G}\mathbf{y}$ *if, and only if,* $\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}$. □

To see this result observe first that if $\mathbf{G}$ is a g-inverse of $\mathbf{A}$, so that $\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}$, then

$$\mathbf{A}\mathbf{x} = \mathbf{y} \implies \mathbf{A}\mathbf{G}\mathbf{A}\mathbf{x} = \mathbf{y} \implies \mathbf{A}\mathbf{G}\mathbf{y} = \mathbf{y} \implies \widetilde{\mathbf{x}} = \mathbf{G}\mathbf{y}$$

is a particular solution to the equation.

An alternative to Theorem 5.7 is the following.

Theorem 5.6 (An Alternate Expression for $\widetilde{\mathbf{x}}$)**.** *When* $\mathbf{A}$ *has* n *columns and* $\mathbf{G}$ *is a g-inverse of* $\mathbf{A}$*, the consistent equations* $\mathbf{A}\mathbf{x} = \mathbf{y}$ *have solutions*

$$\widetilde{\mathbf{x}} = \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{z} \tag{5.4.1}$$

for any arbitrary vector z *of order* n*.* □

Theorem 5.7 (All Possible Solutions)**.** *All solutions of the consistent equations* $\mathbf{A}\mathbf{x} = \mathbf{y}$ *can be generated from either* $\widetilde{\mathbf{x}} = \mathbf{G}\mathbf{y}$ *for* $\mathbf{y} \neq \mathbf{0}$ *by using all possible generalized inverses* $\mathbf{G}$ *of* $\mathbf{A}$ *or from* $\widetilde{\mathbf{x}} = \mathbf{G}\mathbf{y} + (\mathbf{G}\mathbf{A} - \mathbf{I})\mathbf{z}$ *for any specific* $\mathbf{G}$ *by using all possible values of the arbitrary vector* $\mathbf{z}$*.* □

There is one interesting feature of the various solutions that is given in the following theorem:

Theorem 5.8 (Combinations of Solutions)**.** *If* $\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \ldots, \widetilde{\mathbf{x}}_p$*, are any* p *solutions to consistent equations* $\mathbf{A}\mathbf{x} = \mathbf{y}$ *for* $\mathbf{y} \neq \mathbf{0}$ *then* $\overline{\mathbf{x}} = \sum_{j=1}^{p} \omega_j \widetilde{\mathbf{x}}_j$ *is a solution if, and only if,* $\sum_{j=1}^{p} \omega_j = 1$*.* □

We have, to date, had little to say about homogeneous equations $\mathbf{Ax} = \mathbf{0}$? Such sets of equations are always consistent and one possible solution is $\mathbf{x} = \mathbf{0}$. Indeed, if $\mathbf{A}$ has full column rank then $\mathbf{x} = \mathbf{0}$ is the unique solution. However, if $\mathbf{A}$ is of less than full column rank then other solutions are available too and one advantage of Theorem 5.6 over Theorem 5.5 is that it applies even if $\mathbf{y} = \mathbf{0}$. In this case we see that (5.4.1) reduces to

$$\widetilde{\mathbf{x}} = (\mathbf{GA} - \mathbf{I})\mathbf{z}, \tag{5.4.2}$$

where $\mathbf{z}$ remains arbitrary. Of course, a parametric solution of the form (4.6.1) is always available.

One interesting application of homogeneous equations is that the provide an avenue for generating orthogonal solutions. An algorithm for doing this can be described as follows. First find a single solution to the consistent equations $\mathbf{Ax} = \mathbf{0}$ and call it $\widetilde{\mathbf{x}}_1$. By construction this solution will be consistent with the rest of the system. Moreover, we know that it is orthogonal to each of the rows of $\mathbf{A}$. (That is what the original equation is telling us.) So we can augment the original coefficient matrix according to

$$\begin{bmatrix} \mathbf{A} \\ \widetilde{\mathbf{x}}_1' \end{bmatrix} \mathbf{x} = \mathbf{0}.$$

Solving this system of equations will yield a new solution, call it $\widetilde{\mathbf{x}}_2$, say. Not only will $\widetilde{\mathbf{x}}_2$ be a solution to the initial set of equations but, by construction, it will be orthogonal to $\widetilde{\mathbf{x}}_1$. This is a process that can be repeated until you have exhausted the available set of linearly independent solutions. All other solutions must then be linear combinations of the linearly independent set. This raises the question of just how many linearly independent solutions there are? In the case of homogeneous equations $\mathbf{Ax} = \mathbf{0}$ the answer is reasnably intuitively obvious. Recall the discussion of basis sets from Section 1.5. We see that the solution vector $\widetilde{\mathbf{x}}$ is of dimension $n \times 1$ and so it represents a vector in $\mathbb{R}^n$. In order to allow for any possible vector $\mathbf{z}$, which is also of dimension $n \times 1$, we need to find a basis set for $\mathbb{R}^n$, being a set of n linearly independent vectors of dimension $n \times 1$.

5.5 Generalized Inverses and Least Squares

Least squares and regression analysis is a conerstone of econometric analysis. It is, however, both its greatest streangth and its greatest weakness as an inferential tool that least squares has absolutely nothing to do with statistical analysis. Specifically, the existence of a data generating process of any description has absolutely no impact on the ability of least squares to do its thing. That is, of course, because the development of least squares took place in the complete absence of any notion of a data generating process. To emphasize this point one last time, data generating processes are a complete irrelevance to the method of leasts squares. That least squares appears as an optimal outcome in certain statistical models is nothing more than serendipity. Least squares can trace its published history back at least as far as Legendre (1806), which is a supplement to Legendre (1805), although priority remains in question with Gauss claiming to have been using least squares at least a decade earlier than the publication of Legendre's work.[3]

[3]The contributions of Adrien-Marie Legendre (1752–1833) can be explored at https://en.wikipedia.org/wiki/Adrien-Marie_Legendre. For discussion of the question of priority see,

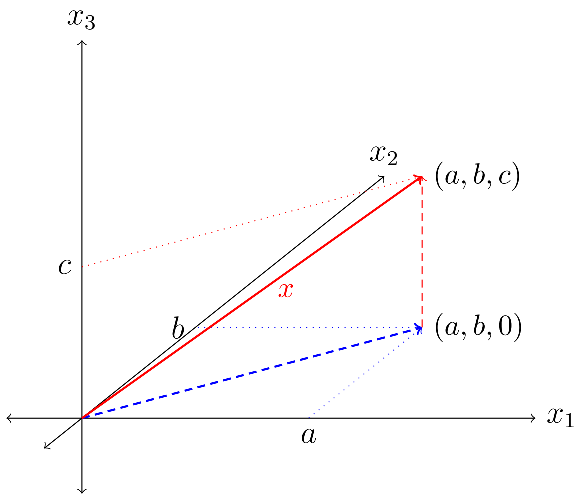


Figure 5.1: Ray from the Origin in 3D

Figure 5.1 depicts the vector $\mathbf{x} = [a, b, c]'$ (the solid red line) in a three dimensional space. By recursively applying Pythagoras Theorem we can work out, first, the length of the dashed blue line joining the origin to the point $(a, b, 0)$, and, second, the length of $\mathbf{x}$. Specifically, the dashed blue line has length $\sqrt{a^2 + b^2}$, from which we deduce that $\mathbf{x}$ has length

$$\sqrt{\left(\sqrt{a^2 + b^2}\right)^2 + c^2} = \sqrt{a^2 + b^2 + c^2}.$$

More generally, if $\mathbf{x} = [x_1, x_2, \ldots, x_k]'$ is a k-vector then it has length

$$\sqrt{\sum_{j=1}^{k} x_j^2} = \sqrt{\mathbf{x}'\mathbf{x}}.$$

This measure of length is known as Euclidean distance.

Now suppose that you have a set of data captured in an n-vector $\mathbf{y}$ that you are trying to explain as a linear combination of a set of k variables for which the data is gathered in a $n \times k$ matrix $\mathbf{X}$. The problem that you face is that of trying to choose an appropriate linear combination of the columns $\mathbf{X}$ to explain what you have observed in $\mathbf{y}$.[4] If we denote an arbitrary linear combination of the columns of $\mathbf{X}$ by $\tilde{y} = \mathbf{X}\boldsymbol{\beta}$, then you might choose to denote your preferred linear combination by $\hat{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$. The question is how best to choose $\hat{\boldsymbol{\beta}}$ from among the candidate set of $\boldsymbol{\beta}$? The least squares approach is to minimize the length of the vector that measures the distance)or residual) between $\mathbf{y}$ and $\tilde{\mathbf{y}}$, namely $\mathbf{e} = \mathbf{y} - \tilde{\mathbf{y}}$. From our discussion of Euclidean distance, we know that length to be $\sqrt{\mathbf{e}'\mathbf{e}}$. However, given that the square root function is monotonic, life become easier if one works with the criterion $S(\boldsymbol{\beta}) = \mathbf{e}'\mathbf{e}$. The least squares solution is then[5]

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} S(\boldsymbol{\beta}).$$

inter alios, Plackett (1972) and Stigler (1981).

[4]Note there is nothing in here that is in any way indicative of a data generating process or a causal relationship of any kind, merely a $\mathbf{y}$ and an $\mathbf{X}$. One may quibble about how $\mathbf{X}$ is chosen but that is not integral to the story being told here.

[5]Recall that $\min_x f(x)$ function gives the minimum value of the function f over all values of

As $S(\boldsymbol{\beta})$ is everywhere continuous we can use techniques of calculus to solve this minimization problem. These techniques are discussed in Section 9 and will be applied here without further discussion. Thus, the first order conditions to find a stationary point are $\partial S(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathbf{0}$ and the second order condition for a minimum is that

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} > \mathbf{0}.$$

The first order conditions yield the set of k equations[6]

$$\begin{aligned}\mathbf{0} = \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})} \\ &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),\end{aligned}$$

from which can be deduced the so-called least squares *normal equations*:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \tag{5.5.1}$$

Note that we have added the '$\hat{\ }$' to $\boldsymbol{\beta}$ once the derivative is equated to zero. These normal equations, or *estimating equations* as they are sometimes called, is just a system of linear equations and we have devoted a lot of time to the study of solving sets of linear equations.

For completeness, we can check the second order condition:

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -2\frac{\partial \mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}'} = 2\mathbf{X}'\mathbf{X}$$

which is indeed positive definite provided that $\mathbf{X}'\mathbf{X}$ is positive definite, which it will be provided that $\mathbf{X}$ has full column rank. In this case we have a unique solution to the normal equations of

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

Given this solution we obtain various other quantities of interest such as

$$\hat{\mathbf{y}} = \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}_X\mathbf{y} \tag{5.5.2a}$$

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{P}_X)\mathbf{y} = \mathbf{M}_X\mathbf{y} \tag{5.5.2b}$$

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{M}_X\mathbf{y})'\mathbf{M}_X\mathbf{y} = \mathbf{y}'\mathbf{M}_X'\mathbf{M}_X\mathbf{y} = \mathbf{y}'\mathbf{M}_X\mathbf{y}, \tag{5.5.2c}$$

as $\mathbf{M}_X' = \mathbf{M}_X$ and $\mathbf{M}_X\mathbf{M}_X = \mathbf{M}_X$; that is, $\mathbf{M}_X$ is symmetric and idempotent, as is $\mathbf{P}_X$. Noting that $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$, we see that

$$\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}\hat{\mathbf{y}} + \hat{\mathbf{e}}'\hat{\mathbf{e}} \implies \hat{\mathbf{y}}\hat{\mathbf{y}} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{y}'\mathbf{P}_X\mathbf{y},$$

because $\hat{\mathbf{y}}'\hat{\mathbf{e}} = \mathbf{y}'\mathbf{P}_X'\mathbf{M}_X\mathbf{y} = \mathbf{y}'\mathbf{P}_X\mathbf{M}_X\mathbf{y} = 0$.

What happens if $\mathbf{X}$ does not have full column rank? First, the matrix $\mathbf{X}'\mathbf{X}$ is only positive semi-definite; that is $\mathbf{X}'\mathbf{X} \geq \mathbf{0}$. This means that any solution to the

x, whereas $\text{argmin}_x\, f(x)$ gives the value of the argument x at which the function f is minimized. Similarly for argmax, although obviously here the discussion is couched in terms of maximization rather than minimization.

[6]Observe that the chain rule given here is slightly different to that given in (v) because there the function F is treated as a scalar-valued function, wheeas here the function $\mathbf{y} - \mathbf{X}\beta$ is a vector-valued function of $\boldsymbol{\beta}$ and so the appropriate chain rule reflects that fact.

normal equations (5.5.1) is not a unique minimzer of the sum of squared residuals. We know this to be true in any event because, second, we know from Theorem 5.6 that there is an infinite number of solutions to (5.5.1) of the form

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} + ((\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} - \mathbf{I}_k)\mathbf{z},$$

where $\mathbf{z}$ is an arbitrary k-vector and $(\mathbf{X}'\mathbf{X})^{-}$ is any g-inverse of $\mathbf{X}'\mathbf{X}$.

Because $\mathbf{X}'\mathbf{X}$ is symmetric it has many properties that are of interest. The following list is taken from Searle (1982, Chapter 8), where proofs can be found, and is but a small subset of those available there.

(i) $\mathbf{X}'\mathbf{X}$ has the full rank decomposition $\mathbf{K}\mathbf{K}'$, where $\mathbf{K}$ has full column rank. Consequently, we can define a Moore-Penrose inverse of $\mathbf{X}'\mathbf{X}$ as

$$(\mathbf{X}'\mathbf{X})^{+} = \mathbf{K}(\mathbf{K}'\mathbf{K})^{-2}\mathbf{K}'.$$

(ii) Let $\mathbf{G} = (\mathbf{X}'\mathbf{X})^{-}$ be any g-inverse of $\mathbf{X}'\mathbf{X}$ then

   (a) $\mathbf{G}'$ is also a g-inverse of $\mathbf{X}'\mathbf{X}$.

   (b) $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}$, so that $\mathbf{G}\mathbf{X}'$ is a g-inverse of $\mathbf{X}$. Similarly, from (i), $\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{X} = \mathbf{X}$ and $\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{X}' = \mathbf{X}'$.

   (c) $\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}\mathbf{G}'\mathbf{X}'$, with both expressions being symmetric, even if $\mathbf{G} \neq \mathbf{G}'$ and regardless of whether or not $\mathbf{G}$ is symmetric.

   (d) $\mathbf{X}\mathbf{G}\mathbf{X}'$ is invariant to $\mathbf{G}$, That is, it doesn't matter which g-inverse appears in the quadratic form it always takes the same value. This is a key result.

As a consequence of these results we can establish the following results for the matrix $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, which are readily extended to $\mathbf{P} = \mathbf{I}_n - \mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$.

(i) $\mathbf{M}$ is symmetric, idempotent and invariant to $(\mathbf{X}'\mathbf{X})^{-}$.

(ii) $\mathbf{M}\mathbf{X} = \mathbf{0}$ and $\mathbf{X}'\mathbf{M} = \mathbf{0}$.

(iii) The rank of $\mathbf{M}$ is $\rho(\mathbf{M}) = n - \rho(\mathbf{X})$.

(iv) Columns of $\mathbf{M}$ are orthogonal to and linearly independent of columns of $\mathbf{X}$.

(v) All n-vectors can be expressed as linear combinations of the columns of $\mathbf{X}$ and $\mathbf{M}$. That is, they span the space $\mathbb{R}^n$ but they are not a basis set, as such, because they comprise a set of $n + k$ vectors, which is larger than required to form a basis set for $\mathbb{R}^n$.

(vi) Any matrix $\mathbf{B}$ for which $\mathbf{B}\mathbf{X} = \mathbf{0}$ has rows that are linear combinations of the rows of $\mathbf{M}$; hence, there always exists some matrix $\mathbf{A}$ such that $\mathbf{B} = \mathbf{A}\mathbf{M}$.

(vii) $\mathbf{M}$ has zero row sums when the row sums of $\mathbf{X}$ are all the same and non-zero.

(viii) $\mathbf{M}$ has zero row sums when a column of $\mathbf{X}$ is $\mathbf{1}_n = [1, 1, \ldots, 1]'$, an n-vector comprised entirely of ones.

(ix) $\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ is a special case of $\mathbf{M}$, with $\mathbf{X} = \mathbf{1}_n$.

(x) If $\mathbf{A}$ is symmetric and satisfies $\mathbf{AX} = \mathbf{0}$, then $\mathbf{A} = \mathbf{MLM}$ for some $\mathbf{L}$.

Given the various results listed above, we see that, even though $\widetilde{\boldsymbol{\beta}}$ is not uniquely defined, quantities such as the residual sum of squares $\mathbf{y}'\mathbf{M}_X\mathbf{y}$ and the explained sum of squares $\mathbf{y}'\mathbf{P}_X\mathbf{y}$ are because they are invariant to the choice of g-inverse. These quantities should still be used with caution, however, it is clear that you cannot generate any result that you may want by simply changing your choice of g-inverse and that there is no return to trying different g-inverses, at least in respect of these quantities.

Chapter 6

Eigenvalues and Eigenvectors

6.1 That Which We Call a Rose By Any Other Word Would Smell as Sweet

Before embarking on an introduction to the concepts of *eigenvalues* and *eigenvectors*, it is worth pointing out that these concepts have other names as well. Specifically, 'eigen' translates from German as 'proper' or 'characteristic' and so eigenvalues and eigenvectors are sometimes known as either *characteristic values* or *characteristic roots* and *characteristic vectors*, respectively. Alternative names also in common use are *latent roots* and *latent vectors*.

6.2 Motivation

The concepts of eignenvectors and eigenvalues arise naturally in a various fields of mathematics and the motivation given below is far from the only one.

Consider a linear equation of the form[1]

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{6.2.1}$$

where we will assume that $\mathbf{A}$ is of dimension $n \times n$ and $\mathbf{x}$ is $n \times 1$. This equation may, for example, describe some production process. In such a process, a set of inputs $\mathbf{x}$, say, are combined according to the coefficients in the matrix $\mathbf{A}$ to produce more units of $\mathbf{x}$ as outputs. Depending on the nature of the process we may have more of $\mathbf{x}$ available at the end of the process than we had at the beginning, in which case $\lambda > 0$, or we may have the same amount, $\lambda = 1$, or we may even have consumed more of $\mathbf{x}$ in the production process than we produced, so that $\lambda < 1$. Indeed, we may have ultimately produced nothing at all or even gone into deficit, so that $\lambda \leq 0$. This kind of idea is closely related to the study of input-output tables considered in Chapter 10.

As it happens, equations such as (6.2.1) arise naturally in a variety of fields of study, including economics, engineering, physics, chemistry, mathematics, and so on. Another example arises in the study of vibrations. In many cases, vibrations tend to be damped so that they decay away at some rate. Examples include things

[1]Note that $\mathbf{A}$ must be square in order for $\mathbf{x}$ to have the same dimensions on either side of the equality.

like plucked guitar strings and automobile suspensions. The way that these things vibrate can be modelled using equations like (6.2.1). In almost all cases the question of interest is to determine whether we can find a vector $\mathbf{x}$ and a constant λ, for a given square matrix $\mathbf{A}$, such that (6.2.1) is satisfied.

If we allow $\mathbf{x} = \mathbf{0}$ then, for all $\mathbf{A}$, $\mathbf{A0} = \mathbf{0}$. For the most part, this observation is totally uninteresting because it tells us nothing about $\mathbf{A}$. The question becomes more interesting if we restrict attention to $\mathbf{x} \neq \mathbf{0}$. In this case, if $\mathbf{A} = \mathbf{I}$ then all $\mathbf{x}$ will satisfy (6.2.1) provided that $\lambda = 1$. Otherwise, the choice of $\mathbf{x}$ and λ becomes more complicated and it is to this problem that we will turn our attention. Our analysis will not preclude the possibility that $\mathbf{A} = \mathbf{I}$, but it will not impose this restriction either and so we will proceed as though $\mathbf{A} \neq \mathbf{I}$.

6.3 The Characteristic Equation of A

Simple re-arrangement of (6.2.1) yields

$$(\mathbf{A} - \lambda \mathbf{I}_n)\mathbf{x} = \mathbf{0}. \tag{6.3.1}$$

The import of this equation is that, if there are any values of $\mathbf{x} \neq \mathbf{0}$ and λ that satisfy (6.2.1), it must be the case that the matrix $\mathbf{A} - \lambda\mathbf{I}$ is singular.[2] (If $\mathbf{A} - \lambda\mathbf{I}$ is non-singular then there is no $\mathbf{x}$ satisfying (6.3.1), which is actually a definition of a non-singular matrix.) Given that $\mathbf{A} - \lambda\mathbf{I}$ is singular, it implies that

$$\det(\mathbf{A} - \lambda \mathbf{I}_n) = 0. \tag{6.3.2}$$

Equation (6.3.2) is called the *characteristic equation* of $\mathbf{A}$.[3] If $\mathbf{A}$ is of dimension $n \times n$ then the determinant is a polynomial of order n in λ. The roots of this polynomial are called the *eigenvalues*, or *characteristic roots*, of $\mathbf{A}$. Once the eigenvalues of $\mathbf{A}$ have been found from (6.3.2) then the corresponding eigenvectors for a given eigenvalue can be solved from (6.3.1). Let us formalize this definition.

Definition 6.3.1. Eigenfunctions of a Matrix.
For any square matrix $\mathbf{A}$ of order n, its eigenvalues are the roots of the characteristic equation (6.3.2), which is a polynomial of order n in λ, and for any given λ the corresponding eigenvectors are any solutions to the equation (6.3.1). □

Note that (6.3.1) is a homogeneous equation and that $\mathbf{A} - \lambda\mathbf{I}_n$ is singular, meaning that it has less than full rank. As we saw in Section 2.4.4, in the case of a singular coefficient matrix, we can find an infinite number of solutions for $\mathbf{x}$. To see this observe that, for any number c,

$$c(\mathbf{A} - \lambda \mathbf{I}_n)\mathbf{x} = c\mathbf{0} = \mathbf{0} \implies (\mathbf{A} - \lambda \mathbf{I}_n)(c\mathbf{x}) = \mathbf{0},$$

[2] Notice that the $\mathbf{x}$ and λ come in pairs; specifically, for any $\mathbf{x}$ satisfying (6.2.1) there will be a corresponding λ. The converse statement is not true, in that there may be multiple $\mathbf{x}$ for any given λ such that (6.2.1) is satisfied. More on this later.

[3] Some authors write this equation as $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$, in essence they have multiplied (6.3.1) by $(-1)^n$, i.e., if $\mathbf{B}$ is an $n \times n$ matrix then $\det(-\mathbf{B}) = (-1)^n \det(\mathbf{B})$; see Property (vii) on page 42. Given that the right-hand side of the equation is zero, it is clear that multiplication by $(-1)^n$ changes nothing. This must also be true then for the left-hand side of the equation. That is, given the equality of the polynomial to zero, multiplication by $(-1)^n$, or indeed any other scale factor, can have no impact on the solution.

and we see immediately that if $\mathbf{x}$ is a characteristic vector of $\mathbf{A}$ then so too is $c\mathbf{x}$ for any number c. We will return to this point later. Before that, let us look at some examples.

Example 6.3.1. Eigenvalues and Eigenvectors of a 2×2 Symmetric Matrix. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix},$$

then the eigenvalues of $\mathbf{A}$ are the solutions to

$$0 = \det(\mathbf{A} - \lambda \mathbf{I}_n) = \begin{vmatrix} 1-\lambda & 2 \\ 2 & 3-\lambda \end{vmatrix} = (1-\lambda)(3-\lambda) - 2\times 2 = \lambda^2 - 4\lambda - 1.$$

We can solve this quadratic in the usual way to obtain the roots $\lambda = 2 \pm \sqrt{5}$.[4] (You should check that both $\lambda = 2+\sqrt{5}$ and $\lambda = 2-\sqrt{5}$ satisfy $\lambda^2 - 4\lambda - 1 = 0$.) Note that if we had, instead, solved $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$ then the characteristic equation reduces to $(\lambda-1)(\lambda-3) - (-2)^2 = \lambda^2 - 4\lambda - 1 = 0$, as before, and nothing is changed.

Having obtained the eigenvalues then we can obtain the corresponding eigenvectors via (6.3.1). When $\lambda_1 = 2+\sqrt{5}$ then we see that

$$(\mathbf{A} - \lambda_1 \mathbf{I}_2)\mathbf{x} = \mathbf{0} \implies \begin{bmatrix} -\left(1+\sqrt{5}\right) & 2 \\ 2 & 1-\sqrt{5} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

If we use elementary row operations to reduce the coefficient matrix to row echelon form then we might first, reduce the 1,1 element to unity and then subtract twice this new first from the second to obtain

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} -\left(1+\sqrt{5}\right)^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -\left(1+\sqrt{5}\right) & 2 \\ 2 & 1-\sqrt{5} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -2/\left(1+\sqrt{5}\right) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 - \dfrac{2x_2}{1+\sqrt{5}} \\ 0 \end{bmatrix}. \end{aligned}$$

And so we have a parametric solution of the form $x_1 = 2x_2/\left(1+\sqrt{5}\right)$, with x_2 arbitrary. That is, corresponding to the eigenvalue λ_1 is the eigenvector $v_1 = [2x_2/\left(1+\sqrt{5}\right), x_2]'$. Similarly, if we use the eigenvalue $\lambda_2 = 2-\sqrt{5}$ then we obtain

$$\begin{bmatrix} x_1 + \dfrac{2x_2}{1-\sqrt{5}} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

so that the corresponding eigenvector is $v_2 = [-2x_2/\left(1-\sqrt{5}\right), x_2]'$, with x_2 arbitrary.

[4] If $ax^2 + bx + c = 0$ then $x = (-b \pm \sqrt{b^2 - 4ac})/(2a)$.

Example 6.3.2. Eigenvalues and Eigenvectors of a 3×3 Asymmetric Matrix.
The characteristic equation for

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 1 & 1 \\ -7 & 2 & -3 \end{bmatrix} \quad \text{is} \quad \begin{vmatrix} 2-\lambda & 2 & 0 \\ 2 & 1-\lambda & 1 \\ -7 & 2 & -3-\lambda \end{vmatrix} = 0.$$

Recalling the diagonal expansion of (3.4.5) and, in particular, (3.4.4) we can immediately write that

$$\det(\mathbf{A} - \lambda \mathbf{I}_3) = (-\lambda)^3 + (-\lambda)^2 \operatorname{tr}(\mathbf{A}) + (-\lambda) \operatorname{tr}_2(\mathbf{A}) + \det(\mathbf{A}),$$

where

$$\begin{aligned} \operatorname{tr}(\mathbf{A}) &= 2 + 1 - 3 = 0 \\ \operatorname{tr}_2(\mathbf{A}) &= \begin{vmatrix} 1 & 1 \\ 2 & -3 \end{vmatrix} + \begin{vmatrix} 2 & 0 \\ -7 & -3 \end{vmatrix} + \begin{vmatrix} 2 & 2 \\ 2 & 1 \end{vmatrix} \\ &= (-3-2) + (-6-0) + (2-4) = -13 \\ \det(\mathbf{A}) &= 2\begin{vmatrix} 1 & 1 \\ 2 & -3 \end{vmatrix} - 2\begin{vmatrix} 2 & 1 \\ -7 & -3 \end{vmatrix} + 0\begin{vmatrix} 2 & 1 \\ -7 & 2 \end{vmatrix} \\ &= 2(-5) - 2(1) = -12, \end{aligned}$$

where I have evaluated $\det(\mathbf{A})$ via a Laplace expansion across the first row. Thus,

$$\det(\mathbf{A} - \lambda \mathbf{I}_3) = -\lambda^3 + 13\lambda - 12,$$

and so

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}_3) = 0 &\implies \lambda^3 - 13\lambda + 12 = 0 \\ \implies (\lambda - 1)(\lambda^2 + \lambda - 12) &= (\lambda - 1)(\lambda - 3)(\lambda + 4) = 0. \end{aligned}$$

That is, the eigenvalues of $\mathbf{A}$ are $\lambda \in \{1, 3, -4\}$. □

It is worth noting that here the eigenvalues of $\mathbf{A}$ are all real but that is not always going to be the case, sometimes they will come in complex conjugate pairs, which are numbers of the form $a \pm b\imath$, where $\imath^2 = -1$ and a and b are both real numbers. For example, the roots of the equation $\lambda^2 + 4 = 0$ are $\lambda = \pm 2\imath$, so here $a = 0$ and $b = 2$.

One useful fact about eigenvalues is the following.

Corollary 6.1 (Diagonal Matrices)**.** *If $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$ is a diagonal matrix of order n then its eigenvalues are the values $d_1, \ldots, d_n$ appearing on its leading diagonal.* □

This result follows immediately from the definition of the determinant of a diagonal matrix, see Section 3.2, Property (vi).

We noted in Section 2.5 that there was a variety of different definitions of rank. Here is yet another one.

Definition 6.3.2. Yet Another Definition of Rank.
The rank of a square matrix is equal to its number of non-zero eigenvalues.

We can illustrate these last two ideas in a single example.

Example 6.3.3. Eigenvalues and Diagonal Matrices.
Consider the matrix $\mathbf{D} = \text{diag}(2, 3, 0)$. We note that such a matrix is sufficiently close to being in reduced row echelon form that we can readily deduce that it is of rank 2. Moreover, because it is a diagonal matrix, by Corollary 6.1, it should be the case that the eigenvalues are 0, 2 and 3, respectively. If this is the case then we have 2 non-zero eigenvalues, which is consistent with Definition 6.3.2. It remains only to confirm that 0, 2, and 3 are the eigenvalus of $\mathbf{D}$. To see this we need to find the roots of the characteristic equation

$$0 = \det(\mathbf{D} - \lambda \mathbf{I}_3) = \begin{vmatrix} 2-\lambda & 0 & 0 \\ 0 & 3-\lambda & 0 \\ 0 & 0 & -\lambda \end{vmatrix} = -\lambda(3-\lambda)(2-\lambda).$$

where the final equality follows from Section 3.2, Property (vi). Clearly, the relevant roots are $\lambda \in \{0, 2, 3\}$, as required. □

One unusual result relating to the characteristic equation of a square matrix is the following:

Theorem 6.1 (Cayley-Hamilton Theorem)**.** *Every matrix satisfies its own characteristic equation. That is, if $p(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I})$ denotes the characteristic polynomial of $\mathbf{A}$, expressed as a function of λ then $p(\mathbf{A}) = 0$.*

Example 6.3.4. The Cayley-Hamilton Theorem.
Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

Then the characterisic equation of $\mathbf{A}$ is a polynomial in λ satisfying

$$p(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}_2) = 0.$$

By (3.4.1), we have immediately

$$p(\lambda) = (-\lambda)^2 + (-\lambda)\,\text{tr}(\mathbf{A}) + \det(\mathbf{A}) = 0 \implies p(\lambda) = \lambda^2 - 5\lambda - 2 = 0.$$

That is, $p(\lambda)$ is weighted sum of the terms λ^2, $\lambda^1 \equiv \lambda$ and $\lambda^0 \equiv 1$. The Cayley-Hamilton Theorem tells us that $p(\mathbf{A}) = 0$, which means that we will be forming a weighted sum of terms $\mathbf{A}^2$, $\mathbf{A}^1 \equiv \mathbf{A}$, and $\mathbf{A}^0 \equiv \mathbf{I}_2$, where the weights are the same as for the corresponding terms in $p(\lambda)$. That is,

$$p(\mathbf{A}) = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - 5 \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - 2\mathbf{I}_2 = \begin{bmatrix} 7-5-2 & 10-10-0 \\ 15-15-0 & 22-20-2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

as required. □

One reason why this result might be of interest is that it provides a recursive way to build up powers of a matrix.

Example 6.3.5. Example 6.3.4 Continued.
Suppose that we wish to find $\mathbf{A}^5$, where $\mathbf{A}$ is as defined in Example 6.3.4. One might consider performing the five multiplications $\mathbf{A} \times \mathbf{A} \times \mathbf{A} \times \mathbf{A} \times \mathbf{A}$. Quicker would be to first calculate $\mathbf{A}^2 = \mathbf{A} \times \mathbf{A}$ and then to calculate $\mathbf{A}^2 \times \mathbf{A}^2 \times \mathbf{A}$, which

only involves 3 matrix multiplications. In the case of a 2×2 matrix, this is not too onerous, but for larger matrices it may become an issue. (Given that you would, presumably, be using a computer for such a calculation, numerical issues are more likely to become a problem than are the physical cost of performing the calculations themselves.) However, even easier is to recognise that in this case $\mathbf{A}^2 = 5\mathbf{A} + 2\mathbf{I}_2$, by the Cayley-Hamilton Theorem, and hence that

$$\begin{aligned}
\mathbf{A}^3 &= \mathbf{A}\mathbf{A}^2 = \mathbf{A}(5\mathbf{A} + 2\mathbf{I}_2) = 5\mathbf{A}^2 + 2\mathbf{A} = 5(5\mathbf{A} + 2\mathbf{I}_2) + 2\mathbf{A} = 27\mathbf{A} + 10\mathbf{I}_2 \\
\mathbf{A}^4 &= \mathbf{A}(27\mathbf{A} + 10\mathbf{I}_2) = 27(5\mathbf{A} + 2\mathbf{I}_2) + 10\mathbf{A} = 145\mathbf{A} + 54\mathbf{I}_2 \\
\mathbf{A}^5 &= \mathbf{A}(145\mathbf{A} + 54\mathbf{I}_2) = 779\mathbf{A} + 290\mathbf{I}_2,
\end{aligned}$$

where we have made repeated use of our expression for $\mathbf{A}^2$. We see that now there is no matrix multiplication to speak of. Rather, two matrices are multiplied by scalars and then there is a matrix addition, which involves far fewer computations then either of the previous suggestions. □

The previous example illustrates some gain in terms of raising matrices to a power, however, there is yet more to be had. Somewhat counter-intuitively, using numbers rather than symbols has made our task more difficult than it need be because it has lost sight of structure in the problem; specifically, the 5 and the 2 in the expression for $\mathbf{A}^2$ are not arbitrary, instead they are functionally related. In particular, $5 = \mathrm{tr}(\mathbf{A}) = \lambda_1 + \lambda_2$ and $2 = -\det \mathbf{A} = -\lambda_1\lambda_2$, where $\lambda_1 \geq \lambda_2$ are the eigenvalues of $\mathbf{A}$. If we use the eigenvalues in our expressions then a remarkable result reveals itself. Starting with the expression for $\mathbf{A}^2$, and defining $\delta = \lambda_2/\lambda_1$, we find

$$\begin{aligned}
\mathbf{A}^2 &= (\lambda_1 + \lambda_2)\mathbf{A} - \lambda_1\lambda_2\mathbf{I}_2 = \lambda_1(1+\delta)\mathbf{A} - \lambda_1\lambda_2\mathbf{I}_2 \\
\mathbf{A}^3 &= (\lambda_1 + \lambda_2)\mathbf{A}^2 - \lambda_1\lambda_2\mathbf{A} = ((\lambda_1+\lambda_2)^2 - \lambda_1\lambda_2)\mathbf{A} - (\lambda_1+\lambda_2)\lambda_1\lambda_2\mathbf{I}_2 \\
&= \lambda_1^2(1+\delta+\delta^2)\mathbf{A} - \lambda_1(1+\delta)\lambda_1\lambda_2\mathbf{I}_2 \\
\mathbf{A}^4 &= \lambda_1^3(1+\delta+\delta^2+\delta^3)\mathbf{A} - \lambda_1^2(1+\delta+\delta^2)\lambda_1\lambda_2\mathbf{I}_2 \\
\mathbf{A}^5 &= \lambda_1^4(1+\delta+\delta^2+\delta^3+\delta^4)\mathbf{A} - \lambda_1^3(1+\delta+\delta^2+\delta^3)\lambda_1\lambda_2\mathbf{I}_2 \\
&\ \ \vdots
\end{aligned}$$

There are two pattern to observe here. First, the coefficient on the identity matrix is always $-\lambda_1\lambda_2$ times the coefficient on $\mathbf{A}$ of the next lower power. Second, the coefficients on $\mathbf{A}$ in the reduction of $\mathbf{A}^k$ take the form of λ_1^{k-1} times the sum of a goemertric progression involving k terms. Recall that the sum of such a progression takes the form

$$1 + \delta + \delta^2 + \cdots + \delta^{k-1} = \begin{cases} \dfrac{1-\delta^k}{1-\delta}, & \text{if } \delta \neq 1, \\ k, & \text{if } \delta = 1. \end{cases}$$

Hence, on noting that $\lambda_1 \geq \lambda_2$ by definition,

$$\lambda_1^{k-1}\sum_{j=0}^{k-1}\delta^j = \begin{cases} \dfrac{\lambda_1^{k-1}(1-\delta^k)}{1-\delta} = \dfrac{(\lambda_1^k - \lambda_2^k)}{\lambda_1 - \lambda_2}, & \lambda_1 > \lambda_2 \\ k\lambda^{k-1}, & \lambda_1 = \lambda_2 = \lambda \text{ (say)}. \end{cases}$$

From this we conclude that

$$\mathbf{A}^k = \begin{cases} \dfrac{(\lambda_1^k - \lambda_2^k)}{\lambda_1 - \lambda_2}\mathbf{A} - \dfrac{\lambda_1\lambda_2(\lambda_1^{k-1} - \lambda_2^{k-1})}{\lambda_1 - \lambda_2}\mathbf{I}_2, & \lambda_1 > \lambda_2 \\ k\lambda^{k-1}\mathbf{A} - (k-1)\lambda^k\mathbf{I}_2, & \lambda_1 = \lambda_2 = \lambda. \end{cases} \tag{6.3.3}$$

Example 6.3.6. Example 6.3.4 Concluded.
The characteristic equation of $\mathbf{A}$ is $\lambda^2 - 5\lambda - 2 = 0$ from which we conclude that the eigenvalues of $\mathbf{A}$ are given by

$$\lambda = \frac{-(-5) \pm \sqrt{(-5)^2 - (4)(-2)}}{2} = 2.5 \pm \sqrt{33/4}.$$

That is, $\lambda_1 \approx 5.3723$ and $\lambda_2 \approx -0.3723$. If we let α_k denote the coefficient on $\mathbf{A}$ in $\mathbf{A}_k$ and β_k denote the coefficient on $\mathbf{I}_2$, so that $A^k = \alpha_k\mathbf{A} + \beta_k\mathbf{I}_2 = \alpha_k\mathbf{A} - \lambda_1\lambda_2\alpha_{k-1}\mathbf{I}_2$, then we have in Table 6.1a the values given by the formulae in (6.3.3) when the eigenvalues differ. In Table 6.1b, we repeat the exercise except this time the matrix

Table 6.1: Table of Coefficients from Equation (6.3.3)

(a) Coefficients Based on $\mathbf{A}$

k	α_k	β_k
1	1	0
2	5	2
3	27	10
4	145	54
5	779	290
6	4148	1558

(b) Coefficients Based on $\mathbf{B} = 2\mathbf{I}_2$

k	α_k	β_k
1	1	0
2	4	-4
3	12	-16
4	32	-48
5	80	-128
6	192	-320

of interest is $\mathbf{B} = 2\mathbf{I}_2$, so that $\lambda = 2$. These calculations are based on that member of (6.3.3) applicable when eigenvalues are equal. □

6.4 Eigenfunctions of Symmetric Matrices

Often we are concerned with the eigenfunctions of symmetric matrices. Symmetric matrices have some important special features which we shall provide here without proof. The interested reader is referred to Searle (1982, Section 11.6).

- The eigenvalues of every real symmetric matrix are real.
- Symmetric matrices are diagonalable. (See the discussion in Section 7.3.)
- The eigenvectors of a symmetric matrix are orthogonal. (Specifically, see the discussion of the spectral decomposition of a symmetric matrix in Section 7.3.2)

6.5 The Elementary Symmetric Functions

As our starting point recall that the characteristic equation of $p \times p$ matrix $\mathbf{A}$, that is $\det(\mathbf{A} - \lambda \mathbf{I}_p) = 0$, is a polynomial of order p in λ, where the coefficients of the polynomial are complicated functions of the elements of $\mathbf{A}$. Let us suppose that this polynomial can be written as

$$f(\lambda) = c_p \lambda^p + c_{p-1}\lambda^{p-1} + \cdots + c_0, \quad c_p \neq 0.$$

One version of the Fundamental Theorem of Algebra is the following.

Theorem 6.2. *Polynomials of order p with complex coefficients have p solutions. The solutions may or may not be distinct. The solutions may be real or complex. Complex solutions appear in complex conjugate pairs, i.e., in the form $x \pm \imath y$, where x and y are real numbers and $\imath^2 = -1$.*

Another version of this result is the following.

Theorem 6.3 (Landau (1965, Theorem 334, p.239))**.** *All polynomials can be written as a product of linear terms:*

$$f(x) = a_p x^p + a_{p-1}x^{p-1} + \cdots + a_0 = a_p \prod_{j=1}^{p}(x - r_j), \quad a_p \neq 0,$$

for suitably defined r_j, $j \in \{0, 1, \ldots, p-1\}$, which are the roots of the equation. If r^* denotes the largest root of f then, for some $x^* \geq r^*$, $f(x)$ will take and stay of the same sign as a_p for all $x > x^*$.

What we need to take out of these last two results is that any univariate polynomial of order p has p roots which may be either real or complex and which may or may not be distinct (some of the r_j in Landau's result may be the same). It transpires that we can relate the roots of a polynomial to its coefficients via a set of results known as Vieta's formulae, which state

$$e_j(r_1, r_2, \ldots, r_p) = (-1)^j \frac{a_{p-j}}{a_p},$$

where

$$\begin{aligned}
e_1(r_1, r_2, \ldots, r_p) &= r_1 + r_2 + \cdots + r_{p-1} + r_p \\
e_2(r_1, r_2, \ldots, r_p) &= (r_1 r_2 + r_1 r_3 + \cdots + r_1 r_p) + (r_2 r_3 + r_2 r_4 + \cdots + r_2 r_p) \\
&\quad + \cdots + r_{p-1} r_p \\
&\vdots \\
e_p(r_1, r_2, \ldots, r_p) &= r_1 r_2 \ldots r_p.
\end{aligned}$$

(See Funkhouser (1930) for a discussion of the historical development of these relations.) The $e_j(r_1, r_2, \ldots, r_p)$, $j = 1, \ldots, p$, are known as the *elementary symmetric functions of the roots* $r_1, r_2, \ldots, r_p$. (By convention we also have $e_0(r_1, r_2, \ldots, r_p) = 1$.) For now it is worth noting that as j increases, $e_j(r_1, r_2, \ldots, r_p)$ becomes an increasingly complicated function of the roots $r_1, r_2, \ldots, r_p$.

Example 6.5.1. Elementary Symmetric Functions.
For $p = 1$:

$$e_1(r_1) = r_1.$$

For $p = 2$:

$$\begin{aligned} e_1(r_1, r_2) &= r_1 + r_2, \\ e_2(r_1, r_2) &= r_1 r_2. \end{aligned}$$

For $p = 3$:

$$\begin{aligned} e_1(r_1, r_2, r_3) &= r_1 + r_2 + r_3, \\ e_2(r_1, r_2, r_3) &= r_1 r_2 + r_1 r_3 + r_2 r_3, \\ e_3(r_1, r_2, r_3) &= r_1 r_2 r_3. \end{aligned}$$

For $p = 4$:

$$\begin{aligned} e_1(r_1, r_2, r_3, r_4) &= r_1 + r_2 + r_3 + r_4, \\ e_2(r_1, r_2, r_3, r_4) &= r_1 r_2 + r_1 r_3 + r_1 r_4 + r_2 r_3 + r_2 r_4 + r_3 r_4, \\ e_3(r_1, r_2, r_3, r_4) &= r_1 r_2 r_3 + r_1 r_2 r_4 + r_1 r_3 r_4 + r_2 r_3 r_4. \\ e_4(r_1, r_2, r_3, r_4) &= r_1 r_2 r_3 r_4. \end{aligned}$$

$\square$

Observe that in the case of *monic polynomials*, where the leading coefficient is equal to unity, i.e., $a_p = 1$ in our context, we have the following identity

$$\begin{aligned} \prod_{j=1}^{p} (x - r_j) = x^p - e_1(r_1, r_2, \ldots, r_p) x^{p-1} + e_2(r_1, r_2, \ldots, r_p) x^{p-2} \\ + \cdots + (-1)^p e_p(r_1, r_2, \ldots, r_p). \end{aligned}$$

As a final aside, note that, if the roots in question came from the characteristic equation of a square matrix, $\mathbf{A}$ say, so that the roots were the characteristic roots or eigenvalues of $\mathbf{A}$, then $e_1(x_1, x_2, \ldots, x_p) = \operatorname{tr} \mathbf{A}$ and $e_p(x_1, x_2, \ldots, x_p) = \det \mathbf{A}$. More broadly, we see that the $e_k(r_1, r_2, \ldots, r_p)$ correspond exactly to the $\operatorname{tr}_k(\mathbf{A})$ defined at (3.4.6).

Chapter 7

Some Useful Matrix Decompositions

7.1 Equivalent Canonical Form

Theorem 7.1 (Equivalent Canonical Form)**.** *Any non-null $m \times n$ matrix $\mathbf{A}$ of rank r is equivalent to*

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{C},$$

say, where $\mathbf{P}$ is an $m \times m$ non-singular matrix and $\mathbf{Q}$ is an $n \times n$ non-singular matrix. □

This theorem says that the matrix $\mathbf{C}$, or $\mathbf{PAQ}$, is equivalent to $\mathbf{A}$ in the sense that they are linked by a set of non-singular transformations. The matrices $\mathbf{P}$ and $\mathbf{Q}$ can be constructed as the product of elementary matrices, so they are matrices capturing a sequence of elementary row operations in the case of $\mathbf{P}$ and a sequence of column operations in the case of $\mathbf{Q}$. The importance of this reduction is that it always exists; that is, we can always find some matrices $\mathbf{P}$ and $\mathbf{Q}$ that satisfy the theorem, although they are not unique; see for example, the discussion at the end of Example 2.2.5.[1] Equally, for any $a \neq 0$, if matrices $\mathbf{P}$ and $\mathbf{Q}$ satisfy the theorem then so too will $\mathbf{P}^* = \alpha\mathbf{P}$ and $\mathbf{Q}^* = \alpha^{-1}\mathbf{Q}$.

We won't provide a separate example at this stage because we have already illustrated the idea in Example 2.2.5 and because Theorem 7.1 is nothing more than the logical extension of the elementary operations that we explored earlier. Moreover, its primary purpose is to provide a foundation for the results that follow.

[1] In Example 2.2.5 we saw that $\mathbf{Q}$ was not unique but there was no corresponding discussion for $\mathbf{P}$. The reason for that is that in Example 2.2.5 there is an implicit normalization in the definition of $\mathbf{P}$, namely that $\mathbf{PA}$ yields the reduced row echelon form. This means that the leading non-zero entry in each row of $\mathbf{PA}$ is unity, which imposes a normalization on the elements of $\mathbf{P}$. In general, there is no such restriction on the definitions of either $\mathbf{P}$ or $\mathbf{Q}$ in obtaining the equivalent canonical form.

7.2 Full Rank Factorization

This is one of the most important factorizations that we will encounter. From Theorem 7.1 we have the equivalent canonical for an $m \times n$ matrix $\mathbf{A}$ of rank r

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0} \end{bmatrix} [\mathbf{I}_r, \mathbf{0}]$$
$$\implies \mathbf{A} = \mathbf{P}^{-1}[\mathbf{I}_r, \mathbf{0}]'[\mathbf{I}_r, \mathbf{0}]\mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{P}^{11} \\ \mathbf{P}^{21} \end{bmatrix} [\mathbf{Q}^{11}\ \mathbf{Q}^{12}] = \mathbf{P}^{-1}_{\cdot 1}\mathbf{Q}^{-1}_{1\cdot},$$

where
$$\mathbf{P}^{-1} = \begin{bmatrix} \mathbf{P}^{11} & \mathbf{P}^{12} \\ \mathbf{P}^{21} & \mathbf{P}^{22} \end{bmatrix} \quad \text{and} \quad \mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{Q}^{11} & \mathbf{Q}^{12} \\ \mathbf{Q}^{21} & \mathbf{Q}^{22} \end{bmatrix}$$
are both partitioned so that the $(1,1)$ block is of dimension $r \times r$. Let $\mathbf{P}^{1:r} = [(\mathbf{P}^{11})', (\mathbf{P}^{21})']'$ be an $m \times r$ matrix comprised of the first r columns of $\mathbf{P}^{-1}$ and $\mathbf{Q}_{1:r} = [\mathbf{Q}^{11}, \mathbf{Q}^{12}]$ be an $r \times n$ matrix comprised of the first r rows of $\mathbf{Q}^{-1}$, $r \leq \min(m,n)$. Because $\mathbf{P}$ and $\mathbf{Q}$ are both non-singular, their rows and columns form sets of linearly independent vectors and so any collection of r of these vectors will have rank r. That is, $\mathbf{P}^{1:r}$ has full column rank and $\mathbf{Q}_{1:r}$ has full row rank. Moreover, because the equivalent canonical form always exists we can always construct a full rank factorization. Let us formalize this result, with a slightly simpler notation, and then look at an example.

Theorem 7.2 (Full Rank Factorization)**.** *Any non-null $m \times n$ matrix $\mathbf{A}$ of rank r can always be factored as $\mathbf{A} = \mathbf{KL}$, where both the $m \times r$ matrix $\mathbf{K}$ and the $r \times n$ matrix $\mathbf{L}$ have rank r.*

Example 7.2.1. A Full Rank Factorization.
Suppose
$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 & 5 \\ 2 & 5 & 1 & 14 \\ 4 & 9 & 3 & 24 \end{bmatrix}.$$

Then, after a little work (use row operations to first reduce $\mathbf{A}$ to row echelon form and then use column operations to sweep out the rows), we see that one possible pair of choices for $\mathbf{P}$ and $\mathbf{Q}$ is

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -2 & -1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{bmatrix} 1 & -2 & -3 & 3 \\ 0 & 1 & 1 & -4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and that
$$\mathbf{C} = \mathbf{PAQ} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Therefore
$$\mathbf{A} = \mathbf{P}^{-1}\mathbf{C}\mathbf{Q}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 5 \\ 0 & 1 & -1 & 4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 5 \\ 0 & 1 & -1 & 4 \end{bmatrix} = \mathbf{KL} \quad \text{(say)}. \qquad \square$$

There is a couple of interesting things to observe in this example.

(i) The inverse of the lower triangular matrix $\mathbf{P}$ is itself lower triangular. Similarly, the inverse of the upper triangular matrix $\mathbf{Q}$ is itself upper triangular. These illustrate general results for triangular matrices (see Property (vii) of Section 1.3.7).

(ii) In exactly the same way as $\mathbf{P}$ and $\mathbf{Q}$ are not unique in the determining the equivalent canonical form, nor are $\mathbf{K}$ and $\mathbf{L}$ in the full rank factorization. Indeed, if $\mathbf{S}$ is an arbitrary non-singular matrix then $\mathbf{A} = \mathbf{K}^*\mathbf{L}^*$, where $\mathbf{K}^* = \mathbf{KS}$ and $\mathbf{L}^* = \mathbf{S}^{-1}\mathbf{L}$. It is not uncommon for uniqueness to be imposed by adopting some normalization, such as all the elements in the first row must be positive or, perhaps, equal to unity. Such normalizations tend to be completely arbitrary, but they are not uncommon in practice.

Example 7.2.2. Moore-Penrose Generalised Inverse.
In Section 5.2 we gave the following definition for a Moore-Penrose generalized inverse of a matrix $\mathbf{A}$:

$$\mathbf{A}^+ = \mathbf{L}'(\mathbf{K}'\mathbf{A}\mathbf{L}')^{-1}\mathbf{K}',$$

where $\mathbf{K}$ and $\mathbf{L}$ are as defined in Theorem 7.2. Using this definition, and the results of Example 7.2.1, if

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 & 5 \\ 2 & 5 & 1 & 14 \\ 4 & 9 & 3 & 24 \end{bmatrix},$$

so that

$$\mathbf{K} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 4 & 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{L} = \begin{bmatrix} 1 & 2 & 1 & 5 \\ 0 & 1 & -1 & 4 \end{bmatrix}$$

then

$$\mathbf{A}^+ = \begin{bmatrix} 1 & 2 & 1 & 5 \\ 0 & 1 & -1 & 4 \end{bmatrix}' \left(\begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 4 & 1 \end{bmatrix}' \begin{bmatrix} 1 & 2 & 1 & 5 \\ 2 & 5 & 1 & 14 \\ 4 & 9 & 3 & 24 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 5 \\ 0 & 1 & -1 & 4 \end{bmatrix}' \right)^{-1} \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 4 & 1 \end{bmatrix}'$$

$$= \frac{1}{702} \begin{bmatrix} 1 & 2 & 1 & 5 \\ 0 & 1 & -1 & 4 \end{bmatrix}' \begin{bmatrix} 162 & -549 \\ -228 & 777 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 4 & 1 \end{bmatrix}'$$

$$= \frac{1}{702} \begin{bmatrix} 162 & -225 & 99 \\ 96 & -129 & 63 \\ 390 & -546 & 234 \\ -102 & 159 & -45 \end{bmatrix}.$$

It is then a simple matter to check the four requirements of a Moore-Penrose inverse; namely (i) that it is indeed a generalized inverse, so that $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, (ii) that is reflexive, so that $\mathbf{A}^{+1}\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$, (iii) that $\mathbf{A}\mathbf{A}^+$ is symmetric, and (iv) that $\mathbf{A}^+\mathbf{A}$

is symmetric. As luck would have it, all these conditions are actually satisfied. You should convince yourself that this is so. (Some computer software adept at matrix calculations, such as Matlab, would be your friend here.) □

Corollary 7.1 (Full Rank Factorization of a Non-Singular Matrix). *When* $\mathbf{A}$ *is non-singular, and hence square* ($m \times m$ *say), then* $\mathbf{PAQ} = \mathbf{I}_m$ *and* $\mathbf{A} = \mathbf{P}^{-1}\mathbf{Q}^{-1}$.

This corollary shows that every non-singular matrix is a product of elementary operators, although this result is implicit in the definition of elementary operators as non-singular matrices.

7.3 Symmetric Matrices

In this section we focus on some special cases of the earlier results that are applicable in the event that the matrix being decomposed is a symmetric matrix.

7.3.1 Congruent Canonical Form of a Symmetric Matrix

The equivalent canonical form of a matrix is such that for an arbitrary matrix $\mathbf{A}$ of rank r, there always exists matrices $\mathbf{P}$ and $\mathbf{Q}$ such that

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

When $\mathbf{A}$ is symmetric, let's call it $\mathbf{S}$ say, then we note that $\mathbf{P}$ and $\mathbf{Q}$ are necessarily of the same dimension, $m \times m$ say, unlike what we saw with the equivalent canonical form of a rectangular array. Moreover, it can be shown that general $\mathbf{Q} = \mathbf{P}^\top$. Let us call this matrix $\mathbf{R}$ to distinguish it from the earlier case. Formally, we have:

Theorem 7.3 (Congruent Canonical Form of a Symmetric Matrix). *For any* $m \times m$ *symmetric matrix* $\mathbf{S}$ *of rank* r *there exists an* $m \times m$ *matrix* $\mathbf{R}$*, say, such that*

$$\mathbf{RSR}^\top = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{7.3.1}$$

In general, $\mathbf{R}$ *may be complex, except for the special case where* $\mathbf{S}$ *is non-negative definite, in which case the elements of* $\mathbf{R}$ *are all real numbers.*

Equation (7.3.1) is known as the *congruent canonical form of a symmetric matrix*, or the *canonical form under congruence of a symmetric matrix*. So we see that it is essentially just a different name for the equivalent canonical form that applies in the case of symmetric matrices. There are two subtleties of which you should be aware. First, in the event that $\mathbf{S}$ is not non-negative definite, then it is possible that $\mathbf{P}$ may contain complex elements. One way around this is work with a variant of the canonical form under congruence if the following way. It is possible to arrange the rows of $\mathbf{R}$ in such a way that

$$\mathbf{TST}^\top = \begin{bmatrix} \mathbf{I}_{r-q} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{7.3.2}$$

where $\mathbf{T}$ denotes the re-arranged version of $\mathbf{R}$. The quantity $r-2q$ is sometimes called the *signature* of $\mathbf{S}$ and the triplet of values $(r-q, q, m-r)$ is sometimes called the *inertia* of $\mathbf{S}$ (Ouellette, 1981).

Second, should one have to use elementary row operations to shuffle rows then care must be taken to preserve symmetry. This is perhaps best illustrated by example. Consider the matrix

$$\mathbf{S} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 20 & 10 \\ 4 & 8 & 10 & 20 \end{bmatrix},$$

and subtract twice the first row from the second, subtract three times the first row from the third, and subtract four times the first row from the fourth row. That is, premultiply $\mathbf{S}$ by the elementary matrix

$$\mathbf{E}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -3 & 0 & 1 & 0 \\ -4 & 0 & 0 & 1 \end{bmatrix}$$

to obtain

$$\mathbf{E}_1\mathbf{S} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 11 & -2 \\ 0 & 0 & -2 & 4 \end{bmatrix}.$$

There are two interesting features of $\mathbf{E}_1\mathbf{S}$. First, the second row is comprised entirely of zeros and what we have learned about transformations to reduced row echelon form is that null rows should be moved below non-null rows. Second, the lower right block of the matrix, $\left[\begin{smallmatrix} 11 & -2 \\ -2 & 4 \end{smallmatrix}\right]$, has retained its symmetry. If we were to simply interchange the second and fourth rows then this would be lost. In particular,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 11 & -2 \\ 0 & 0 & -2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & -2 & 4 \\ 0 & 0 & 11 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and we see that the symmetry is lost in the block that is now $\left[\begin{smallmatrix} -2 & 4 \\ 11 & -22 \end{smallmatrix}\right]$. Instead, it is necessary to swap the second row to become the fourth, whilst simultaneously shifting the bottom two row up a row, maintaining their position relative to one another. Thus,

$$\mathbf{E}_2\mathbf{E}_1\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 11 & -2 \\ 0 & 0 & -2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 11 & -2 \\ 0 & 0 & -2 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

maintains the symmetry of the block in question. Adding two elevenths of the second row to the third yields

$$\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{2}{11} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 11 & -2 \\ 0 & 0 & -2 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 11 & -2 \\ 0 & 0 & 0 & \frac{40}{11} \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Finally, scaling the matrix appropriately gives us the reduced row echelon form:

$$\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{11} & 0 & 0 \\ 0 & 0 & \frac{11}{40} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 11 & -2 \\ 0 & 0 & 0 & \frac{40}{11} \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & -\frac{2}{11} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

We are not quite there because, as we shall see, the non-zero off-diagonal elements remaining are going to cause a slight problem. If we for the quadratic form $\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{S}\mathbf{E}_1'\mathbf{E}_2'\mathbf{E}_3'\mathbf{E}_4'$ then we obtain

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{11} & 0 & 0 \\ 0 & 0 & \frac{11}{40} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{7.3.3}$$

and so we need a further scaling to reach the canonical form under congruence. That is,

$$\mathbf{E}_5 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{11} & 0 & 0 \\ 0 & 0 & \sqrt{40/11} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then

$$\mathbf{R} = \mathbf{E}_5\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{3}{\sqrt{11}} & 0 & \frac{1}{\sqrt{11}} & 0 \\ -\frac{25}{\sqrt{110}} & 0 & \frac{1}{\sqrt{110}} & \sqrt{\frac{11}{40}} \\ -2 & 1 & 0 & 0 \end{bmatrix}.$$

Equation (7.3.3) is an example of what is known as the *diagonal form of a symmetric matrix*, which we define as follows.

Theorem 7.4 (Diagonal Form of a Symmetric Matrix)**.** *For any $m \times m$ symmetric matrix $\mathbf{S}$ of rank r there exists an $m \times m$ matrix $\mathbf{W}$, say, such that*

$$\mathbf{W}\mathbf{S}\mathbf{W}^\top = \begin{bmatrix} \mathbf{D}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{D}_r$ is an $r \times r$ diagonal matrix of full rank.

Corollary 7.2. *If $\mathbf{S}$ is an $m \times m$ non-negative definite symmetric matrix of rank r then $\mathbf{R} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{R}$ is as as defined in Theorem 7.3.1,*

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_r^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-r} \end{bmatrix},$$

and both $\mathbf{D}_r$ and $\mathbf{W}$ are as defined in Theorem 7.4.

We note that the source of complex numbers in $\mathbf{R}$ arises from the scaling by $\mathbf{D}_r^{-1/2}$, because it is only when $\mathbf{S}$ is non-negative definite that $\mathbf{D}_r$ is comprised entirely of positive numbers, otherwise it may contain negative numbers.

7.3.2 The Spectral Decomposition of a Symmetric Matrix

If $\mathbf{S}$ is a symmetric matrix of order m, with eigenvalues $\lambda_1, \ldots, \lambda_m$, and corresponding orthonormal eigenvectors $\mathbf{h}_1, \ldots, \mathbf{h}_m$, then

$$\mathbf{S} = \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}',$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_m)$ and $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_m]$.[2] To see this, observe that $\mathbf{H}\mathbf{H}' = \mathbf{I}_m$ on account of the columns of $\mathbf{H}$ being orthonormal. But

$$\mathbf{H}\mathbf{H}' = [\mathbf{h}_1, \ldots, \mathbf{h}_m] \begin{bmatrix} \mathbf{h}_1' \\ \vdots \\ \mathbf{h}_m' \end{bmatrix} = \sum_{j=1}^{m} \mathbf{h}_j \mathbf{h}_j'.$$

Therefore,

$$\mathbf{S} = \mathbf{S}\mathbf{I}_m = \mathbf{S}\mathbf{H}\mathbf{H}' = \mathbf{S}\sum_{j=1}^{m} \mathbf{h}_j\mathbf{h}_j' = \sum_{j=1}^{m} \mathbf{S}\mathbf{h}_j\mathbf{h}_j'.$$

But, by definition, $\mathbf{S}\mathbf{h}_j = \lambda_j \mathbf{h}_j$, and so

$$\mathbf{S} = \sum_{j=1}^{m} \lambda_j \mathbf{h}_j \mathbf{h}_j = \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}',$$

where the final equality follows on noting that

$$\mathbf{H}\boldsymbol{\Lambda} = [\lambda_1 \mathbf{h}_1, \ldots, \lambda_m \mathbf{h}_m],$$

so that

$$\mathbf{H}\boldsymbol{\Lambda}\mathbf{H}' = [\lambda_1 \mathbf{h}_1, \ldots, \lambda_m \mathbf{h}_m] \begin{bmatrix} \mathbf{h}_1' \\ \vdots \\ \mathbf{h}_m' \end{bmatrix} = \sum_{j=1}^{m} \lambda_j \mathbf{h}_j \mathbf{h}_j'.$$

(That $\mathbf{H}\mathbf{H}' = \sum_{j=1}^{m} \mathbf{h}_j\mathbf{h}_j'$ follows on setting $\lambda_j = 1$ for all $j = 1, \ldots, m$.)

We can run exactly the same arguments for powers of $\mathbf{S}$. For example, suppose that k is a positive integer, then

$$\begin{aligned}\mathbf{S}^k &= \mathbf{S}^k \sum_{j=1}^{m} \mathbf{h}_j\mathbf{h}_j' = \mathbf{S}^{k-1} \sum_{j=1}^{m} \mathbf{S}\mathbf{h}_j\mathbf{h}_j' = \mathbf{S}^{k-1} \sum_{j=1}^{m} \lambda_j \mathbf{h}_j\mathbf{h}_j' = \mathbf{S}^{k-2} \sum_{j=1}^{m} \lambda_j \mathbf{S}\mathbf{h}_j\mathbf{h}_j' = \cdots \\ &= \sum_{j=1}^{m} \lambda_j^k \mathbf{h}_j\mathbf{h}_j' = \mathbf{H}\boldsymbol{\Lambda}^k\mathbf{H}'.\end{aligned}$$

[2] As we saw in Chapter 6, eigenvectors are not uniquely defined. In search of uniqueness we often impose various restrictions on the eigenvectors. One common set of restrictions is that the eigenvectors be orthonormal, so that

$$\mathbf{h}_j'\mathbf{h}_k = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise.} \end{cases}$$

By implication, we then have $\mathbf{H}'\mathbf{H} = \mathbf{H}\mathbf{H}' = \mathbf{I}_m$. Even this is insufficient for uniqueness as both $\mathbf{H}$ and $-\mathbf{H}$ will satisfy this restriction. So, it is common to further require that the first element of each eigenvector be positive, or equivalently that the elements in the first row of $\mathbf{H}$ all be positive.

If $\mathbf{S}$ is non-singular then the same argument works for k that are negative integers. Once you are convinced that this is so then the step to fractional k is straightforward. In particular, $k = \pm\frac{1}{2}$ gives us square roots of $\mathbf{S}$ and $\mathbf{S}^{-1}$ in the form of symmetric matrices. This is in contrast to alternative definitions in terms of triangular matrices or a full rank decomposition. Sometimes the symmetry matters.

Chapter 8

Kronecker Products and Vectorization

This section is concerned with properties of various matrix operators that commonly occur together, especially in multivariate analysis. For many more properties than those given here, the interested reader is referred to Lütkepohl (1996).

8.1 Kronecker Product

Let $\mathbf{A}$ denote an $m \times n$ matrix with typical element $a_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, and let $\mathbf{B}$ denote a $p \times q$ matrix. Then the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$, denoted $\mathbf{A} \otimes \mathbf{B}$, is the $mp \times nq$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \ldots & a_{1,n}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \ldots & a_{2,n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}\mathbf{B} & a_{m,2}\mathbf{B} & \ldots & a_{m,n}\mathbf{B} \end{bmatrix}.$$

Some useful properties of Kronecker products include:

(i) $\alpha\mathbf{A} \otimes \beta\mathbf{B} = \alpha\beta(\mathbf{A} \otimes \mathbf{B})$, for α, β scalar.

(ii) $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$.

(iii) $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$.

(iv) $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$.

(v) $\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = \operatorname{tr}(\mathbf{A}) \times \operatorname{tr}(\mathbf{B})$, for $\mathbf{A}$, $\mathbf{B}$ both square.

(vi) $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$.

(vii) $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.

(viii) $\det(\mathbf{A} \otimes \mathbf{B}) = \det(\mathbf{A})^n \det(\mathbf{B})^m$, for $\mathbf{A}$ $m \times m$ and $\mathbf{B}$ $n \times n$ (so both $\mathbf{A}$ and $\mathbf{B}$ must be square).

8.2 Vectorization: vec and vech

The process of vectorization involves taking the columns of a matrix and stacking them into a single vector. This process is represented by the vec operator. For example, if the $m \times 1$ vector $\mathbf{a}_j$ is the jth vector of an $m \times n$ matrix $\mathbf{A}$ then

$$\text{vec}(A) = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}$$

is an $mn \times 1$ vector.

Vectorization and Kronecker products play nicely together and often occur together. Two key properties are:

(i) $\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A}) \, \text{vec}(\mathbf{B})$, for matrices $\mathbf{A}(m \times n)$, $\mathbf{B}(n \times p)$, $\mathbf{C}(p \times q)$.

(ii) If, additionally, $\mathbf{D}(q \times m)$:

$$\begin{aligned} \text{tr}(\mathbf{ABCD}) &= (\text{vec}(\mathbf{D}'))'(\mathbf{C}' \otimes \mathbf{A}) \, \text{vec}(\mathbf{B}) \\ &= (\text{vec}(\mathbf{A}'))'(\mathbf{D}' \otimes \mathbf{B}) \, \text{vec}(\mathbf{C}) \\ &= (\text{vec}(\mathbf{B}'))'(\mathbf{A}' \otimes \mathbf{C}) \, \text{vec}(\mathbf{D}) \\ &= (\text{vec}(\mathbf{C}'))'(\mathbf{B}' \otimes \mathbf{D}) \, \text{vec}(\mathbf{A}). \end{aligned}$$

From these we can derive a variety of useful forms, including:

(iii) $\text{tr}(\mathbf{ABC}) = (\text{vec}(\mathbf{A}'))'(\mathbf{I}_m \otimes \mathbf{B}) \, \text{vec}(\mathbf{C})$.

(iv) $\text{tr}(\mathbf{AX}'\mathbf{BXC}) = (\text{vec}(\mathbf{X}))'(\mathbf{A}'\mathbf{C}' \otimes \mathbf{B}) \, \text{vec}(\mathbf{X}) = (\text{vec}(\mathbf{X}))'(\mathbf{CA} \otimes \mathbf{B}') \, \text{vec}(\mathbf{X})$.

(v) Without the trace operator we have:

$$\begin{aligned} (\text{vec}(\mathbf{D}'))'(\mathbf{C}' \otimes \mathbf{A}) \, \text{vec}(\mathbf{B}) &= (\text{vec}(\mathbf{D}'))' \, \text{vec}(\mathbf{ABC}) \\ &= (\text{vec}(\mathbf{A}'\mathbf{D}'))' \, \text{vec}(\mathbf{BC}) \\ &= (\text{vec}(\mathbf{A}'\mathbf{D}'\mathbf{C}'))' \, \text{vec}(\mathbf{B}) \end{aligned}$$

There is one specialisation of the notion of vectorisation that is useful when working with symmetric matrices. If $\mathbf{S}$ denotes a $p \times p$ symmetric matrix then it contains only $p(p+1)/2$ distinct elements, comprised of those on the leading diagonal on those below it, with those above the leading diagonal being identical to those below. The vech, or *half-vectorization*, operator functions in the same way as does the vec operator except that for each column of its matrix argument it only keeps those elements on or below the leading diagonal. Its name comes from the notion of half-vectorisation. For example, suppose that

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix}.$$

Then we see that

$$\text{vec}(\mathbf{S}) = \begin{bmatrix} s_{11} & s_{21} & s_{31} & s_{12} & s_{22} & s_{32} & s_{13} & s_{23} & s_{33} \end{bmatrix}'$$

whereas

$$\text{vech}(\mathbf{S}) = \begin{bmatrix} s_{11} & s_{21} & s_{31} & s_{22} & s_{32} & s_{33} \end{bmatrix}'.$$

Noting that $s_{12} = s_{21}$, $s_{13} = s_{31}$, and $s_{23} = s_{32}$, by the symmetry of $\mathbf{S}$, we see that the use of the vec operator in this context potentially introduces some unnecessary redundancy that the vech operator ignores.

Because, in the context of symmetric matrices $\mathbf{S}$, the elements of vec($\mathbf{S}$) are the same as those of vech($\mathbf{S}$) with some repetitions, there must exist matrices $\mathbf{D}_p$ and $\mathbf{H}$ such that

$$\text{vec}(\mathbf{S}) = \mathbf{D}_p \,\text{vech}(\mathbf{S}). \qquad \text{and} \qquad \text{vech}(\mathbf{S}) = \mathbf{H}\,\text{vec}(\mathbf{S})$$

These matrices, $\mathbf{D}_p$ and $\mathbf{H}$, are called the duplication and elimination matrices, respectively. Continuing our example, we see that

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$\mathbf{D_p} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

It can be shown that these matrices have the following properties:

- The duplication matrix $\mathbf{D}_p$ is unique, of full column rank $p(p+1)/2$, with p^2 rows, so that $\mathbf{D}_p'\mathbf{D}_p$ is non-singular. For example,

$$\mathbf{D}_p'\mathbf{D}_p = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

  which is evidently non-singular.

- $\mathbf{D}_p$ is non-square, and hence has no inverse, but its Moore-Penrose inverse takes the form $\mathbf{D}_p^+ = (\mathbf{D}_p'\mathbf{D}_p)^{-1}\mathbf{D}_p'$.

- $\mathbf{H}$ is not unique, but does have full row rank $p(p+1)/2$, with p^2 columns. For example, and alternative choice for $\mathbf{H}$ in the previous example is

$$\mathbf{H}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

  Given that there are 3 duplicated elements in $\mathbf{S}$, we see that there are 3! possible choice of $\mathbf{H}$ available. In general, if the symmetric matrix is of dimension $p \times p$ then there will be $p(p-1)/2$ possible choices of $\mathbf{H}$.

- $\mathbf{H}$ is a left inverse of $\mathbf{G}$, i.e., $\mathbf{HD}_p = \mathbf{I}_{p(p+1)/2}$. It follows that $\mathbf{D}_p^+$ is one possible form of the elimination matrix. Completing the previous example, we see that $\mathbf{S}$ is of dimension 3×3, so that $p = 3$, and

$$\mathbf{HD}_p = \mathbf{I}_6,$$

  as required.

For a more complete discussion of both the vec and the vech operators see Searle (1982, Section 12.9) and the references cited therein.

Chapter 9

Matrix Calculus: Derivatives and Differentials

9.1 Some Fundamentals and Notation

We could begin at the beginning and define both derivatives and partial derivatives as limits and take things from there; that is,

(i) If $f(x)$ is some scalar-valued differentiable function of some variable x then we can define the *derivative* of f with respect to x, denoted either $f'(x)$ or $\mathrm{d}\, f(x)/\, \mathrm{d}\, x \equiv \mathrm{d}\, f/\, \mathrm{d}\, x$, as

$$\frac{\mathrm{d}\, f(x)}{\mathrm{d}\, x} \equiv f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

(ii) If $f(x_1, \ldots, x_n)$ is some scalar-valued differentiable function of some set of variables $x_1, \ldots, x_n$ then we can define the *partial derivative* of f with respect to any one member of the set, x_j say, denoted either $f_j'(x)$ or $\partial f(x_1, \ldots, x_n)/\partial x_j \equiv \partial f/\partial x_j$, as

$$\begin{aligned}\frac{\partial f(x_1, \ldots, x_n)}{\partial x_j} &\equiv f_j'(x_1, \ldots, x_n) \\ &= \lim_{h\to 0} \frac{f(x_1, \ldots, x_j + h, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}.\end{aligned}$$

The advantage of the $'$ notation is that it is strongly suggestive of the fact that the (partial) derivatives so obtained remain functions of the original set of variables. (Sometimes they may only be constant functions, but that is okay.) This allows us to differentiate these functions too. So, for example, one may differentiate a second time to obtain second order derivatives, denoted

$$\frac{\mathrm{d}^2\, f(x)}{\mathrm{d}\, x^2} \qquad \text{or} \qquad f''(x),$$

or second order partial derivatives, denoted

$$\frac{\partial^2 f(x_1, \ldots, x_n)}{\partial x_j^2} \qquad \text{or} \qquad f_j''(x_1, \ldots, x_n).$$

Of course, with partial derivatives you may not always want to differentiate with respect to the same variable all the time and so we have the notion of cross-partial derivatives, something like

$$\frac{\partial^2 f(x_1,\dots,x_n)}{\partial x_i \partial x_j} \qquad \text{or} \qquad f''_{i,j}(x_1,\dots,x_n).$$

The disadvantage of the $'$ notation is that it quickly becomes cumbersome if you differentiate more than twice. Typically a slight variant is adopted whereby derivatives of order higher than (say) two or three are denoted $f^{(k)}(x_1,\dots,x_n)$, where k is the order of differentiation and subscripts can be added as required. Another variant is to drop the superscript altogether, as the number of subscripts will indicate the order of differentiation. We note Clairaut's Theorem which states, in essence, that given sufficient continuity the order of (partial) differentiation does not matter. (Here order relates to the ordering in which each partial derivative is taken, i.e., which variable do you differentiate with respect to first, which is second, and so on, rather than the number of times the original function is differentiated.) For example, second-order partial derivatives are symmetric:

$$\frac{\partial^2 f(x_1,\dots,x_n)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x_1,\dots,x_n)}{\partial x_j \partial x_i}.$$

Once one starts thinking about multivariable functions then one may wish to know the impact of all the the inputs changing simultaneously, which brings us to the notion of a total differential, which takes the form

$$\begin{aligned} \mathrm{d}\, f(x_1,\dots,x_n) = \frac{\partial f(x_1,\dots,x_n)}{\partial x_1}\,\mathrm{d}\, x_1 + \frac{\partial f(x_1,\dots,x_n)}{\partial x_2}\,\mathrm{d}\, x_2 + \dots \\ \dots + \frac{\partial f(x_1,\dots,x_n)}{\partial x_n}\,\mathrm{d}\, x_n \end{aligned}$$

or, simply,

$$\mathrm{d}\, f = \frac{\partial f}{\partial x_1}\,\mathrm{d}\, x_1 + \frac{\partial f}{\partial x_2}\,\mathrm{d}\, x_2 + \dots + \frac{\partial f}{\partial x_n}\,\mathrm{d}\, x_n. \tag{9.1.1}$$

As I said, we could do all that, but I will assume that you already know it as there seems little point in starting the study of calculus with the matrix case. This little discursion did, however, allow us to establish notation. In what follows, I will express derivatives in the fractional notation that was originally due to Leibnitz,[1] rather than the $'$ notation of Lagrange.[2] On balance, we are probably better off saving the $'$ notation to denote matrix transpose, although I can't guarantee that this will always be the case, so make sure that you know what a $'$ is denoting in any given circumstance.

[1]Gottfried Wilhelm von Leibnitz (or Leibniz) (1648–1716) was a great German mathematician, historian, and philosopher. Among his many, many achievements, Leibnitz is credited with contemporaneously inventing ideas of differential and integral calculus independently of Isaac Newton who, in the British tradition is typically given precedence. For a discussion of the issue see, for example https://en.wikipedia.org/wiki/Leibniz%E2%80%93Newton_calculus_controversy. Regardless of the merits of either side, Newton's notation is largely abandoned in favour of that of Leibnitz.

[2]Joseph-Louis Lagrange (1736–1813) was a great Italian mathematician who made significant contributions in many fields of mathematics, who spent significant portions of his life working in first Germany and later France.

9.2 Notational Conventions

There are three types of situations that we are going to have to deal with; namely, scalar-valued functions, vector-valued functions, and matrix-valued functions, that we might denote $f(\cdot)$, $\mathbf{f}(\cdot)$, and $\mathbf{F}(\cdot)$, respectively. These three classes of functions may have scalar arguments, vector arguments, or matrix arguments. For example, $y = x^2$ is a scalar-valued function of a scalar argument. We will hereafter ignore this case because it is simply that of introductory courses on calculus. The determinant or trace of a square matrix are examples of scalar-valued functions of matrix argument. We see that $[x, y]' = [r\cos\theta, r\sin\theta]'$ is an example of a vector-valued function of the vector of arguments $[r, \theta]'$. Finally, $\mathbf{X}^{-1}$ is an example of matrix function of a matrix argument.

If the function that we seek todifferentiate is scalar-valued then matrix calculus is reasonably straight-forward. In particular, for a scalar-valued function of a vector such as $y = f(x_1, \ldots, x_n)$, we shall define

$$\frac{\partial y}{\partial \mathbf{x}} = \left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \ldots, \frac{\partial y}{\partial x_n}\right]',$$

a column vector. Similarly, for the scalar-valued function of a matrix $y = f(\mathbf{X})$ it is sometimes convenient to express the resultant as a matrix:

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y_1}{\partial x_{1n}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2mn}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{m1}} & \frac{\partial y}{\partial x_{m2}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix}. \tag{9.2.1}$$

The question is then how to deal with vector-valued or matrix-valued functions? We will adopt the convention that we are always working with column vectors. So, for example, vector-valued functions will be defined so that $\mathbf{f}(\cdot)$ is a column vector. In the case of a matrix-valued, function we will work with either $\text{vec}(\mathbf{F}(\cdot))$ or $\text{vech}(\mathbf{F}(\cdot))$ in the event that $\mathbf{F}()$ is symmetric, with the vectors so defined being column vectors.

Moving on, if we have a vector-valued function $\mathbf{y} = \mathbf{f}(\mathbf{x})$ which is a vector of functions of the vector of arguments $\mathbf{x}$, as defined in our first convention, we see that if $\mathbf{y} = [y_1, \ldots, y_m]'$ is $m \times 1$ and $\mathbf{x} = [x_1, \ldots, x_n]'$ is $n \times 1$ then the *Jacobian*, or *gradient*, matrix is the $n \times m$ matrix of partial derivatives

$$\mathbf{J} = \frac{\partial \mathbf{y}'}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Note, if using the gradient terminology then it is common to use the symbol nabla (∇) to denote the differentiation process; that is, to write $\nabla \mathbf{y}$ or $\nabla_{\mathbf{x}} \mathbf{y}$ rather than $\partial \mathbf{y}'/\partial \mathbf{x}$. The case of (9.2.1) notwithstanding, as it can be handled within this convention, this convention reduces the problem of multivariable differentiation to that of differentiating a vector with respect to another vector, which is a relatively straight-forward problem to deal with.

One advantage of this notation is that it allows us to extend to the problem of second order derivatives of scalar-valued function, so-called *hessian matrices*, *viz.*

$$\frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}'} = \frac{\partial}{\partial \mathbf{x}} \left(\frac{\partial y}{\partial \mathbf{x}} \right)' = \frac{\partial}{\partial \mathbf{x}} \frac{\partial y}{\partial \mathbf{x}'} = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1^2} & \frac{\partial^2 y}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \frac{\partial^2 y}{\partial x_1 \partial x_2} & \frac{\partial^2 y}{\partial x_2^2} & \cdots & \frac{\partial^2 y}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_1 \partial x_n} & \frac{\partial^2 y}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 y}{\partial x_n^2} \end{bmatrix},$$

which is actually an application of (9.2.1).

9.3 Some Well-Known Results

There are lots of places where one can find results on matrix calculus. For example, pretty much all texts on multivariate analysis Muirhead (1982); Gupta and Nagar (2000); Mardia, Kent, and Bibby (1979), included in this number are many econometrics texts, especially older econometrics texts where matrix algebra was more prevalent (Amemiya, 1985; Maddala, 1992; Magnus, 2017; Pollock, 1979; Ruud, 2000; Theil, 1971), and there are many mathematics texts on multivariable calculus (Apostol, 1969; Courant and John, 1974; Moskowitz and Paliogiannis, 2010) to name but a few. There are also books on matrix algebra written by econometricians for econometricians, including Abadir and Magnus (2005); Lütkepohl (1996) and Magnus and Neudecker (1988). Below we list some of the more commonly encountered results. In what follows, unless otherwise defined, we shall let $\mathbf{c}$ denote an $n \times 1$ constant vector, so not a function of anything that we are differentiating with respect to, $\mathbf{A}$ a square constant matrix of appropriate dimension, and $\mathbf{B}$ an $m \times n$ constant matrix.

(i) $\dfrac{\partial \mathbf{x}'\mathbf{c}}{\partial \mathbf{x}} = \mathbf{c}.$ □

This is a scalar-valued function differentiated with respect to a (column) vector. We see that $\mathbf{x}'\mathbf{c} = \sum_{j=1}^n c_j x_j$ and so

$$\frac{\partial \mathbf{c}'\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{c}'\mathbf{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \mathbf{c}'\mathbf{x}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \mathbf{c},$$

as required.

(ii) $\dfrac{\partial \mathbf{x}'\mathbf{B}}{\partial \mathbf{x}} = \mathbf{B}.$ □

This is an example of a row vector differentiated with respect to a column vector. As $\mathbf{x}'\mathbf{B}$ is $1 \times n$ and $\mathbf{x}$ is $m \times 1$, we would expect to obtain a matrix of derivatives of order $m \times n$. If we write $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$, so that $\mathbf{b}_j$ denotes the jth column of $\mathbf{B}$, then we see that each element of the $1 \times n$ vector $\mathbf{x}'\mathbf{B}$ is of the form $\mathbf{x}'\mathbf{b}_j$ for $j = 1, \ldots, n$. From (i) we see that $\partial \mathbf{x}'\mathbf{b}_j / \partial \mathbf{x} = \mathbf{b}_j$. Thus,

$$\frac{\partial \mathbf{x}'\mathbf{B}}{\partial \mathbf{x}} = \left[\tfrac{\partial \mathbf{x}'\mathbf{b}_1}{\partial \mathbf{x}}, \ldots, \tfrac{\partial \mathbf{x}'\mathbf{b}_n}{\partial \mathbf{x}} \right] = \left[\mathbf{b}_1, \ldots, \mathbf{b}_n \right] = \mathbf{B},$$

as required.

(iii) Next we consider a quadratic form:

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \begin{cases} 2\mathbf{A}\mathbf{x}, & \text{if } \mathbf{A} \text{ is symmetric,} \\ (\mathbf{A}' + \mathbf{A})\mathbf{x}, & \text{otherwise.} \end{cases} \qquad \square$$

Again we have a scalar-valued function differentiated with respect to a vector, which we will return as a column vector. Observe that

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_j x_k.$$

In order to visualise what comes next it will be helpful to lay out the terms of this double sum in a two-dimensional array:

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \begin{array}{lllll} a_{11}x_1^2 & a_{12}x_1x_2 & + a_{13}x_1x_3 & + \cdots & + a_{1n}x_1x_n \\ + a_{21}x_1x_2 + & a_{22}x_2^2 & + a_{23}x_2x_3 & + \cdots & + a_{2n}x_2x_n \\ + a_{31}x_1x_3 & + a_{32}x_2x_3 + & a_{33}x_3^2 & + \cdots & + a_{3n}x_3x_n \\ \vdots & \vdots & & \ddots & \vdots \\ + a_{n1}x_1x_n & + a_{n2}x_2x_n & + a_{n3}x_3x_n & + \cdots + & a_{nn}x_n^2, \end{array}$$

where, in an obvious notation, a_{jk} denotes the jkth element of $\mathbf{A}$. Moving forward, we are going to break this sum up into three distinct sums: (i) the sum of terms on the leading diagonal, (ii) the sum of terms below the leading diagonal, what we will call the sub-diagonal terms, and (iii) the terms above the leading diagonal, the super-diagonal terms. That is, we gather the terms thus:

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \underbrace{\sum_{j=1}^{n} a_{jj}x_j^2}_{\text{elements on leading diagonal}} + \underbrace{\sum_{j=2}^{n}\sum_{k=1}^{j-1} a_{jk}x_jx_k}_{\text{sum of sub-diagonal elements}} + \underbrace{\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} a_{jk}x_jx_k}_{\text{sum of super-diagonal elements}},$$

Each of these three terms will be partially differentiated with respect to each of the elements of $\mathbf{x}$. In the case of the terms on the leading diagonal the differentiation is straight-forward. With the other two sums, care must be taken to ensure that all terms are gathered. For example, among the sub-diagonal terms x_1 only occurs among the $n-1$ terms in the first column of the array. In the case of x_2, there are the $n-2$ terms in the second column of the array and also the term in the second row of the first column giving a total of $n-1$ terms to differentiate. Similarly, if we go looking to terms involving x_3 then we find not only the $n-3$ terms in the third column but also the two terms in the first two columns of the third row. Again we have $n-1$ terms, involving cross-products of x_3 with all the other elements in $\mathbf{x}$, but not the square of x_3 with itself. We see the transpose of the same pattern in the super-diagonal elements. Hence,

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_n} \end{bmatrix}$$

$$
= \begin{bmatrix}
2a_{11}x_1 + \left(\sum_{j=2}^n a_{j1}x_j\right) + \left(\sum_{j=2}^n a_{1j}x_j\right) \\
2a_{22}x_2 + \left(a_{21}x_1 + \sum_{j=3}^n a_{j2}x_j\right) + \left(a_{12}x_1 + \sum_{j=3}^n a_{2j}x_j\right) \\
2a_{33}x_3 + \left(\sum_{j=1}^2 a_{3j}x_j + \sum_{j=4}^n a_{j3}x_j\right) + \left(\sum_{j=1}^2 a_{j3}x_j + \sum_{j=4}^n a_{3j}x_j\right) \\
\vdots \\
2a_{kk}x_k + \left(\sum_{j=1}^{k-1} a_{kj}x_j + \sum_{j=k+1}^n a_{jk}x_j\right) + \left(\sum_{j=1}^{k-1} a_{jk}x_j + \sum_{j=k+1}^n a_{kj}x_j\right) \\
\vdots \\
2a_{nn}x_n + \left(\sum_{j=1}^{n-1} a_{nj}x_j\right) + \left(\sum_{j=1}^{n-1} a_{jn}x_j\right)
\end{bmatrix}
$$

$$
= \begin{bmatrix}
\sum_{j=1}^n a_{j1}x_j + \sum_{j=1}^n a_{1j}x_j \\
\left(\sum_{j=1}^2 a_{2j}x_j + \sum_{j=3}^n a_{j2}x_j\right) + \left(\sum_{j=1}^2 a_{j2}x_j + \sum_{j=3}^n a_{2j}x_j\right) \\
\left(\sum_{j=1}^3 a_{3j}x_j + \sum_{j=4}^n a_{j3}x_j\right) + \left(\sum_{j=1}^3 a_{j3}x_j + \sum_{j=4}^n a_{3j}x_j\right) \\
\vdots \\
\left(\sum_{j=1}^k a_{kj}x_j + \sum_{j=k+1}^n a_{jk}x_j\right) + \left(\sum_{j=1}^k a_{jk}x_j + \sum_{j=k+1}^n a_{kj}x_j\right) \\
\vdots \\
\left(\sum_{j=1}^n a_{nj}x_j\right) + \left(\sum_{j=1}^n a_{jn}x_j\right)
\end{bmatrix}
$$

$$
= \begin{bmatrix}
\sum_{j=1}^n a_{j1}x_j + \sum_{j=1}^n a_{1j}x_j \\
\sum_{j=1}^n a_{j2}x_j + \sum_{j=1}^n a_{2j}x_j \\
\sum_{j=1}^n a_{j3}x_j + \sum_{j=1}^n a_{3j}x_j \\
\vdots \\
\sum_{j=1}^n a_{jk}x_j + \sum_{j=1}^n a_{kj}x_j \\
\vdots \\
\sum_{j=1}^n a_{nj}x_j + \sum_{j=1}^n a_{jn}x_j
\end{bmatrix}
$$

$$
= \mathbf{A}'\mathbf{x} + \mathbf{A}\mathbf{x} = \begin{cases} 2\mathbf{A}\mathbf{x}, & \text{if } \mathbf{A} \text{ is symmetric,} \\ (\mathbf{A}' + \mathbf{A})\mathbf{x}, & \text{otherwise,} \end{cases}
$$

as required.

(iv) Let $\mathbf{A}$ denote a symmetric $m \times m$ matrix that is not a function of the elements of the $m \times n$ matrix $\mathbf{B}$. Then

$$
\frac{\partial \mathbf{B}'\mathbf{A}\mathbf{B}}{\partial \mathbf{B}} = \mathbf{G},
$$

where $\mathbf{G}$ is of dimension $n^2 \times mn$ and is defined below. □

To begin, observe that

$$
\mathbf{B}'\mathbf{A}\mathbf{B} = \begin{bmatrix}
\mathbf{b}_1'\mathbf{A}\mathbf{b}_1 & \mathbf{b}_1'\mathbf{A}\mathbf{b}_2 & \dots & \mathbf{b}_1'\mathbf{A}\mathbf{b}_n \\
\mathbf{b}_2'\mathbf{A}\mathbf{b}_1 & \mathbf{b}_2'\mathbf{A}\mathbf{b}_2 & \dots & \mathbf{b}_2'\mathbf{A}\mathbf{b}_n \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{b}_n'\mathbf{A}\mathbf{b}_1 & \mathbf{b}_n'\mathbf{A}\mathbf{b}_2 & \dots & \mathbf{b}_n'\mathbf{A}\mathbf{b}_n
\end{bmatrix},
$$

where $\mathbf{b}_j$, $j = 1, \ldots, n$ denotes the jth column of $\mathbf{B}$, i.e., $\mathbf{b}_j$ is an $m \times 1$ vector. So

$$\operatorname{vec}(\mathbf{B}'\mathbf{A}\mathbf{B}) = [\mathbf{b}_1'\mathbf{A}\mathbf{b}_1, \mathbf{b}_2'\mathbf{A}\mathbf{b}_1, \ldots, \mathbf{b}_n'\mathbf{A}\mathbf{b}_1, \mathbf{b}_1'\mathbf{A}\mathbf{b}_2, \ldots, \mathbf{b}_n'\mathbf{A}\mathbf{b}_n]'$$

and

$$\operatorname{vec}(\mathbf{B}) = [\mathbf{b}_1', \ldots, \mathbf{b}_n']'.$$

We see that there will be three types of derivatives encountered:

$$\frac{\partial \mathbf{b}_i'\mathbf{A}\mathbf{b}_j}{\partial \mathbf{b}_k} = \begin{cases} 2\mathbf{A}\mathbf{b}_j, & i = j = k, \\ \mathbf{A}\mathbf{b}_j & i = k \neq j, \\ \mathbf{0}, & i \neq k, j \neq k, \end{cases}$$

where we have exploited the symmetry of $\mathbf{A}$ and our previous results: (i) and (iii).[3] Thus,

$$\mathbf{G} = \begin{bmatrix} 2\mathbf{b}_1'\mathbf{A} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{b}_2'\mathbf{A} & \mathbf{b}_1'\mathbf{A} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{b}_3'\mathbf{A} & \mathbf{0} & \mathbf{b}_1'\mathbf{A} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{b}_n'\mathbf{A} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{b}_1'\mathbf{A} \\ \mathbf{b}_2'\mathbf{A} & \mathbf{b}_1'\mathbf{A} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & 2\mathbf{b}_2'\mathbf{A} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_3'\mathbf{A} & \mathbf{b}_2'\mathbf{A} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_4'\mathbf{A} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & 2\mathbf{b}_n'\mathbf{A} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n \otimes \mathbf{b}_1'\mathbf{A} \\ \mathbf{I}_n \otimes \mathbf{b}_2'\mathbf{A} \\ \vdots \\ \mathbf{I}_n \otimes \mathbf{b}_n'\mathbf{A} \end{bmatrix} + \mathbf{I}_n \otimes \mathbf{B}'\mathbf{A}.$$

(v) For $m \times m$ $\mathbf{X}$, with distinct elements x_{ij}, $\dfrac{\partial \operatorname{tr}(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{I}$. □

Differentiating a scalar function with respect to the elements of a matrix. The trace is the sum of the diagonal elements of $\mathbf{X}$. Differentiation of this sum with respect to any off-diagonal elements of $\mathbf{X}$ must yield zero. Differentiation with respect to any diagonal element must yield unity.

(vi) Let $\mathbf{A}$ be $m \times n$ and let $\mathbf{X}$ be $n \times m$, so that $\mathbf{AX}$ is $m \times m$. Then

$$\frac{\partial \operatorname{tr}(\mathbf{AX})}{\partial \mathbf{X}} = \mathbf{A}'. \qquad \square$$

As $\operatorname{tr}(\mathbf{AX}) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} x_{ji}$ we see that $\dfrac{\mathrm{d} \operatorname{tr}(\mathbf{AX})}{\mathrm{d}\, x_{ij}} = a_{ji}$ and the result follows.

[3] Note that, because $\mathbf{b}_i'\mathbf{A}\mathbf{b}_j = \mathbf{b}_j'\mathbf{A}\mathbf{b}_i$, if $i \neq k$ but $j = k$ then

$$\frac{\partial \mathbf{b}_i'\mathbf{A}\mathbf{b}_j}{\partial \mathbf{b}_k} = \mathbf{A}\mathbf{b}_i.$$

(vii) For $m \times m$ $\mathbf{X}$, $\dfrac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{C} = \det(\mathbf{X})(\mathbf{X}')^{-1}$, where $\mathbf{C}$ denotes the cofactor matrix of $\mathbf{X}$. The second equality is only valid if $\mathbf{X}$ is non-singular. □

We are again differentialing a scalar-valued function with respect to a matrix and so we can express the result in the form

$$\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial \det(\mathbf{X})}{\partial x_{11}} & \frac{\partial \det(\mathbf{X})}{\partial x_{12}} & \cdots & \frac{\partial \det(\mathbf{X})}{\partial x_{1m}} \\ \frac{\partial \det(\mathbf{X})}{\partial x_{21}} & \frac{\partial \det(\mathbf{X})}{\partial x_{22}} & \cdots & \frac{\partial \det(\mathbf{X})}{\partial x_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \det(\mathbf{X})}{\partial x_{m1}} & \frac{\partial \det(\mathbf{X})}{\partial x_{m2}} & \cdots & \frac{\partial \det(\mathbf{X})}{\partial x_{mm}} \end{bmatrix}. \tag{9.3.1}$$

Recall from Section 3.3 that, in order to calculate the determinant of a matrix, we can expand across any row, instance, according to

$$\det(\mathbf{X}) = \sum_{k=1}^{m} x_{jk} c_{jk}, \tag{9.3.2}$$

where c_{jk} denotes the cofactor of x_{jk}. Of particular importance is the fact that c_{jk} is not a function of x_{jk}. Thus, for all the elements in the jth row of (9.3.1), we can use the corresponding version of (9.3.2) to find that the derivative of $\det \mathbf{X}$ with respect to the jkth element of $\mathbf{X}$ is simply c_{jk}. The final result follows from the definition of the inverse matrix in terms of the adjoint (or transpose of the cofactor) matrix.

(viii) If $\mathbf{X}^{-1}$ is an $m \times m$ nonsingular symmetric matrix, with $\mathbf{A}$ $n \times m$ and $\mathbf{B}$ $m \times p$ then

$$\frac{\partial(\operatorname{vec}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}))'}{\partial \operatorname{vech}(\mathbf{X})} = -\mathbf{D}_p'(\mathbf{X}^{-1}\mathbf{B} \otimes (\mathbf{X}^{-1})'\mathbf{A}'),$$

where $\mathbf{D}_m$ is as defined in Section 8.2. Alternatively, if $\mathbf{X}$ is asymmetric then

$$\frac{\partial(\operatorname{vec}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}))'}{\partial \operatorname{vec}(\mathbf{X})} = -\mathbf{X}^{-1}\mathbf{B} \otimes (\mathbf{X}^{-1})'\mathbf{A}' \qquad \square$$

This is much more complicated than any of our earlier results but it is instructive to think how one might tackle the problem. First, imagine that the elements of $\mathbf{X}$ are functions of some scalar quantity z, say. Then denote the process of differentiating each element of $\mathbf{X}$ with respect to z by

$$\frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}.$$

Now, consider differentiating the identity $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}_m$ with respect to z:

$$\begin{aligned} \frac{\mathrm{d}\,\mathbf{X}\mathbf{X}^{-1}}{\mathrm{d}\,z} &= \frac{\mathrm{d}\,\mathbf{I}_m}{\mathrm{d}\,z} \\ \Longrightarrow \frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}\mathbf{X}^{-1} + \mathbf{X}\frac{\mathrm{d}\,\mathbf{X}^{-1}}{\mathrm{d}\,z} &= \mathbf{0} \\ \Longrightarrow \mathbf{X}\frac{\mathrm{d}\,\mathbf{X}^{-1}}{\mathrm{d}\,z} &= -\frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}\mathbf{X}^{-1} \end{aligned}$$

$$\Longrightarrow \frac{\mathrm{d}\,\mathbf{X}^{-1}}{\mathrm{d}\,z} = -\mathbf{X}^{-1}\frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}\mathbf{X}^{-1}.$$

From Section 8.2, Property (i), $\mathrm{vec}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}) = (\mathbf{B}' \otimes \mathbf{A})\,\mathrm{vec}(\mathbf{X}^{-1})$ and so, assuming that neither $\mathbf{A}$ nor $\mathbf{B}$ have elements that are functions of z, we see that

$$\begin{aligned}
\frac{\mathrm{d}(\mathrm{vec}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}))'}{\mathrm{d}\,z} &= \frac{\mathrm{d}\,\mathrm{vec}(\mathbf{X}^{-1})'}{\mathrm{d}\,z}(\mathbf{B}\otimes\mathbf{A}') \\
&= \left(\mathrm{vec}\left(\frac{\mathrm{d}\,\mathbf{X}^{-1}}{\mathrm{d}\,z}\right)\right)'(\mathbf{B}\otimes\mathbf{A}') \\
&= -\left(\mathrm{vec}\left(\mathbf{X}^{-1}\frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}\mathbf{X}^{-1}\right)\right)'(\mathbf{B}\otimes\mathbf{A}')
\end{aligned}$$

where the second equality uses Property (iv) of Section 8.1. If we once again exapnd the vec of the product we obtain

$$\begin{aligned}
\frac{\mathrm{d}(\mathrm{vec}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}))'}{\mathrm{d}\,z} &= -\left(\mathrm{vec}\left(\frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}\right)\right)'(\mathbf{X}^{-1}\otimes(\mathbf{X}^{-1})')(\mathbf{B}\otimes\mathbf{A}') \\
&= -\left(\mathrm{vec}\left(\frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}\right)\right)'(\mathbf{X}^{-1}\mathbf{B}\otimes(\mathbf{X}^{-1})'\mathbf{A}').
\end{aligned}$$

where this most recent equality exploits Section 8.2, Property (vi). How we proceed from here depends upon whether or not $\mathbf{X}$ is symmetric. The simplest case is where it is not. The result is then completed on allowing the z to be any element of $\mathrm{vec}(\mathbf{X})$. As

$$\left(\mathrm{vec}\left(\frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}\right)\right)' = \frac{\mathrm{d}\,(\mathrm{vec}\,(\mathbf{X}))'}{\mathrm{d}\,z},$$

we see that

$$\begin{aligned}
\frac{\partial(\mathrm{vec}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}))'}{\partial\,\mathrm{vec}(\mathbf{X})} &= -\frac{\partial(\mathrm{vec}(\mathbf{X}))'}{\partial\,\mathrm{vec}(\mathbf{X})}(\mathbf{X}^{-1}\mathbf{B}\otimes(\mathbf{X}^{-1})'\mathbf{A}') \\
&= -\mathbf{I}_{m^2}(\mathbf{X}^{-1}\mathbf{B}\otimes(\mathbf{X}^{-1})'\mathbf{A}') \\
&= -\mathbf{X}^{-1}\mathbf{B}\otimes(\mathbf{X}^{-1})'\mathbf{A}'
\end{aligned}$$

as required.

If, on the other hand, $\mathbf{X}$ is symmetric then we need to take that symmetry into account. To see this, consider a simple 2×2 matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

Suppose that $\mathbf{X}$ is asymmetric, so that $x_{12} \neq x_{21}$. Then

$$\begin{aligned}
\frac{\partial(\mathrm{vec}(\mathbf{X}))'}{\partial\,\mathrm{vec}(\mathbf{X})} &= \frac{\partial[x_{11}, x_{21}, x_{12}, x_{22}]}{\partial[x_{11}, x_{21}, x_{12}, x_{22}]'} \\
&= \begin{bmatrix}
\frac{\mathrm{d}\,x_{11}}{\mathrm{d}\,x_{11}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{11}} & \frac{\mathrm{d}\,x_{12}}{\mathrm{d}\,x_{11}} & \frac{\mathrm{d}\,x_{22}}{\mathrm{d}\,x_{11}} \\
\frac{\mathrm{d}\,x_{11}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{12}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{22}}{\mathrm{d}\,x_{21}} \\
\frac{\mathrm{d}\,x_{11}}{\mathrm{d}\,x_{12}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{12}} & \frac{\mathrm{d}\,x_{12}}{\mathrm{d}\,x_{12}} & \frac{\mathrm{d}\,x_{22}}{\mathrm{d}\,x_{12}} \\
\frac{\mathrm{d}\,x_{11}}{\mathrm{d}\,x_{22}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{22}} & \frac{\mathrm{d}\,x_{12}}{\mathrm{d}\,x_{22}} & \frac{\mathrm{d}\,x_{22}}{\mathrm{d}\,x_{22}}
\end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \mathbf{I}_{2^2} = \mathbf{I}_4.
\end{aligned}$$

However, if $\mathbf{X}$ is symmetric, so that $x_{12} = x_{21}$ then

$$\frac{\partial(\mathrm{vec}(\mathbf{X}))'}{\partial\,\mathrm{vec}(\mathbf{X})} = \frac{\partial[x_{11}, x_{21}, x_{21}, x_{22}]}{\partial[x_{11}, x_{21}, x_{21}, x_{22}]'}$$

$$= \begin{bmatrix} \frac{\mathrm{d}\,x_{11}}{\mathrm{d}\,x_{11}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{11}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{11}} & \frac{\mathrm{d}\,x_{22}}{\mathrm{d}\,x_{11}} \\ \frac{\mathrm{d}\,x_{11}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{22}}{\mathrm{d}\,x_{21}} \\ \frac{\mathrm{d}\,x_{11}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{21}} & \frac{\mathrm{d}\,x_{22}}{\mathrm{d}\,x_{21}} \\ \frac{\mathrm{d}\,x_{11}}{\mathrm{d}\,x_{22}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{22}} & \frac{\mathrm{d}\,x_{21}}{\mathrm{d}\,x_{22}} & \frac{\mathrm{d}\,x_{22}}{\mathrm{d}\,x_{22}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \neq \mathbf{I}_4.$$

Indeed, in this latter case, we see that if we write $\mathrm{vec}(\mathbf{X}) = \mathbf{D}_2\,\mathrm{vech}(\mathbf{X})$, where the 4×3 duplication matrix

$$\mathbf{D}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

then the expression[4]

$$\mathbf{D}_2 \frac{\partial\,\mathrm{vech}(\mathbf{X})'}{\partial\,\mathrm{vech}(\mathbf{X})} = \mathbf{D}_2\mathbf{I}_3 = \mathbf{D}_2.$$

Combining these results, we obtain

$$\begin{aligned} \frac{\mathrm{d}(\mathrm{vec}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}))'}{\mathrm{d}\,z} &= -\left(\mathrm{vec}\left(\frac{\mathrm{d}\,\mathbf{X}}{\mathrm{d}\,z}\right)\right)'(\mathbf{X}^{-1}\mathbf{B} \otimes (\mathbf{X}^{-1})'\mathbf{A}') \\ &= -\left(\frac{\partial\,\mathrm{vech}(\mathbf{X})'}{\partial\,\mathrm{vech}(\mathbf{X})}\right)\mathbf{D}_p'(\mathbf{X}^{-1}\mathbf{B} \otimes (\mathbf{X}^{-1})'\mathbf{A}') \\ &= -\mathbf{D}_p'(\mathbf{X}^{-1}\mathbf{B} \otimes (\mathbf{X}^{-1})'\mathbf{A}'). \end{aligned}$$

Note that we have inadvertently provided a definition for the duplication matrix, namely

$$\frac{\partial\,\mathrm{vec}(\mathbf{S})}{\partial\,\mathrm{vech}(\mathbf{S})} = \mathbf{D}_p \implies \partial\,\mathrm{vec}(\mathbf{S}) = \mathbf{D}_p \partial\,\mathrm{vech}(\mathbf{S}) \implies \mathrm{vec}(\mathbf{S}) = \mathbf{D}_p\,\mathrm{vech}(\mathbf{S}),$$

for any $p \times p$ symmetric matrix $\mathbf{S}$.

9.4 Standard Rules of Differentiation

There are a number of rules of differentiation that you will have learned in the context of scalar-valued functions of scalar arguments that can be ported across more or less directly to the matrix case. Sometimes, however, care must be taken in terms of the order with which certain terms appear.

[4] More generally, for $\mathbf{X}$ $p \times p$ and symmetric,

$$\frac{\partial\,\mathrm{vech}(\mathbf{X})'}{\partial\,\mathrm{vech}(\mathbf{X})} = \mathbf{I}_{p(p+1)/2}.$$

9.4.1 Scalar-Valued Functions of Matrix Argument

Below are some rules for differentiating real valued fubctions with matrix argument.

(i) If $\mathbf{X}$ is $m \times n$ and $c \in \mathbb{R}$ then $\dfrac{\partial c}{\partial \mathbf{X}} = \mathbf{0}$.

(ii) (Linearity) If $\mathbf{X}$ is $m \times n$, and $f(\mathbf{X})$ and $g(\mathbf{X})$ are real-valued scalar functions, with $c_1, c_2 \in \mathbb{R}$, then

$$\frac{\partial[c_1 f(\mathbf{X}) + c_2 g(\mathbf{X})]}{\partial \mathbf{X}} = c_1 \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} + c_2 \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}}.$$

(iii) (Product Rule) If $\mathbf{X}$ is $m \times n$, and $f(\mathbf{X})$ and $g(\mathbf{X})$ are real-valued scalar functions then

$$\frac{\partial f(\mathbf{X}) g(\mathbf{X})}{\partial \mathbf{X}} = f(\mathbf{X}) \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}} + g(\mathbf{X}) \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}.$$

(iv) (Ratio or Quotient Rule) If $\mathbf{X}$ is $m \times n$, and $f(\mathbf{X})$ and $g(\mathbf{X}) \neq 0$ are real-valued scalar functions then

$$\frac{\partial f(\mathbf{X})/g(\mathbf{X})}{\partial \mathbf{X}} = \frac{1}{(g(\mathbf{X}))^2} \left[g(\mathbf{X}) \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} - f(\mathbf{X}) \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}} \right].$$

(v) (Chain Rule) If $\mathbf{X}$ is $m \times n$, and $y = f(\mathbf{X})$ and $g(y)$ are real-valued scalar functions then

$$\frac{\partial g(f(\mathbf{X}))}{\partial \mathbf{X}} = \frac{\partial g(y)}{\partial y} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}.$$

We will leave examples of these few rules for another time. Be mindful of the fact that there are many, many more such rules that we might have chosen to present. Again, Lütkepohl (1996, Chapter 10) is a good source of these.

9.5 Differentials

Recall the formula for a total differential given in (9.1.1) that stated, for a function $f \equiv f(x_1, \ldots, x_n)$,

$$\mathrm{d}\, f = \frac{\partial f}{\partial x_1} \mathrm{d}\, x_1 + \frac{\partial f}{\partial x_2} \mathrm{d}\, x_2 + \ldots + \frac{\partial f}{\partial x_n} \mathrm{d}\, x_n.$$

Given our earlier developments we might write this in matrix notation as

$$\mathrm{d}\, f = \frac{\partial f}{\partial \mathbf{x}'} \mathbf{dx},$$

where $\mathbf{dx} = [\mathrm{d}\, x_1, \ldots, \mathrm{d}\, x_n]$.

Suppose now that we have a set of m such functions $y_1(x_1, \ldots, x_n), \ldots, y_m(x_1, \ldots, x_n)$, so that

$$\mathrm{d}\, y_1(x_1, \ldots, x_n) = \frac{\partial y_1}{\partial x_1} \mathrm{d}\, x_1 + \frac{\partial y_1}{\partial x_2} \mathrm{d}\, x_2 + \ldots + \frac{\partial y_1}{\partial x_n} \mathrm{d}\, x_n$$

$$\vdots$$

$$\mathrm{d}\, y_m(x_1,\ldots,x_n) = \frac{\partial y_m}{\partial x_1}\,\mathrm{d}\, x_1 + \frac{\partial y_m}{\partial x_2}\,\mathrm{d}\, x_2 + \ldots + \frac{\partial y_m}{\partial x_n}\,\mathrm{d}\, x_n$$

or, in an obvious extension of our earlier matrix notation.

$$\mathbf{dy} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}'}\,\mathbf{dx}.$$

The terms $\mathbf{dx}$ and $\mathbf{dy}$ are termed differentials and we note that that they are linear functions of one another, which often makes working with them relatively simple. Moreoever, we note that the weighting matrix in the relationship is none other that the Jacobian matrix that has been occupying our time of late. That is, if a set of variables $\mathbf{y}$ is related to another set of variables $\mathbf{x}$, and if we can find a linear relationship between the two corresponding sets of differentials then we have found the Jacobian matrix.

One reason differentials have proved popular in this context stems from some relatively simple rules for working with them. The following draws heavily on Magnus and Neudecker (1988, Chapters 8 & 9). Adopting their notation we have for u and v real-valued differentiable functions, with α a real constant:

$$\begin{aligned}
\mathbf{d}\,\alpha &= 0, \\
\mathbf{d}(\alpha u) &= \alpha\,\mathbf{d}\,u, \\
\mathbf{d}(u+v) &= \mathbf{d}\,u + \mathbf{d}\,v, \\
\mathbf{d}(u-v) &= \mathbf{d}\,u - \mathbf{d}\,v, \\
\mathbf{d}(uv) &= (\mathbf{d}\,u)v + u(\mathbf{d}\,v), \\
\mathbf{d}\left(\frac{u}{v}\right) &= \frac{v\,\mathbf{d}\,u - u\,\mathbf{d}\,v}{v^2} \qquad\qquad (v \neq 0).
\end{aligned}$$

Common results include:[5]

$$\begin{aligned}
\mathbf{d}\,u^\alpha &= \alpha u^{\alpha-1}\,\mathbf{d}\,u, \\
\mathbf{d}\log u &= u^{-1}\,\mathbf{d}\,u \qquad\qquad (u > 0), \\
\mathbf{d}\,e^u &= e^u\,\mathbf{d}\,u, \\
\mathbf{d}\,\alpha^u &= \alpha^u \log\alpha\,\mathbf{d}\,u \qquad\qquad (\alpha > 0).
\end{aligned}$$

Similar results hold if,in additional to our earlier definitions, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ are matrix functions and $\mathbf{A}$ is a matrix of real constants:

$$\begin{aligned}
\mathbf{d}\,\mathbf{A} &= 0, \\
\mathbf{d}(\alpha\mathbf{U}) &= \alpha\,\mathbf{d}\,\mathbf{U}, \\
\mathbf{d}(\mathbf{U}+\mathbf{V}) &= \mathbf{d}\,\mathbf{U} + \mathbf{d}\,\mathbf{V}, \\
\mathbf{d}(\mathbf{U}-\mathbf{V}) &= \mathbf{d}\,\mathbf{U} - \mathbf{d}\,\mathbf{V},
\end{aligned}$$

[5] The domain of definition of the power function u^α depends on the nature of α. If α is a positive integer then u^α is defined for all real u; but if α is a negative integer or zero, the point $u = 0$ must be excluded. If α is a rational fraction, e.g. $\alpha = p/q$ (where p and q are integers and we can always assume that $q > 0$), then $u^\alpha = \sqrt[q]{u^p}$, so that the function is determined for all values of u when q is odd, and only for $u \geq 0$ when q is even. In cases where α is irrational, the function is defined for $u > 0$.

$$\mathbf{d}(\mathbf{UVW}) = (\mathbf{d}\,\mathbf{U})\mathbf{VW} + \mathbf{U}(\mathbf{d}\,\mathbf{V})\mathbf{W} + \mathbf{UV}(\mathbf{d}\,\mathbf{W}).$$

Finally, we have some special results,

$$\begin{aligned} \mathbf{d}(\mathbf{U}\otimes\mathbf{V}) &= (\mathbf{d}\,\mathbf{U})\otimes\mathbf{V} + \mathbf{U}\otimes\mathbf{d}\,\mathbf{V}, \\ \mathbf{d}\,\mathbf{U}' &= (\mathbf{d}\,\mathbf{U})', \\ \mathbf{d}\,\mathrm{vec}(\mathbf{U}) &= \mathrm{vec}(\mathbf{d}\,\mathbf{U}), \\ \mathbf{d}\,\mathrm{tr}\{\mathbf{U}\} &= \mathrm{tr}\{\mathbf{d}\,\mathbf{U}\}, \\ \mathbf{d}\,\mathrm{tr}\{\mathbf{X}'\mathbf{X}\} &= 2\,\mathrm{tr}\{\mathbf{X}'\,\mathbf{d}\,\mathbf{X}\}. \end{aligned}$$

Armed only with these results we can re-visit some of our earlier results and consider some new results.

Example 9.5.1. Differential of a Quadratic Form.
Differentials

$$\mathbf{d}\,\mathbf{x}'\mathbf{A}\mathbf{x} = (\mathbf{d}\,\mathbf{x})'\mathbf{A}\mathbf{x} + \mathbf{x}'\mathbf{A}\,\mathbf{d}\,\mathbf{x} = ((\mathbf{d}\,\mathbf{x})'\mathbf{A}\mathbf{x})' + \mathbf{x}'\mathbf{A}\,\mathbf{d}\,\mathbf{x} = \mathbf{x}'(\mathbf{A}'+\mathbf{A})\,\mathbf{d}\,\mathbf{x},$$

which implies that $\mathbf{J} = \partial\mathbf{x}'\mathbf{A}\mathbf{x}/\partial\mathbf{x}' = \mathbf{x}'(\mathbf{A}'+\mathbf{A})$, as we saw earlier.[6]

Example 9.5.2. Differential of a Trace.
First, recall our earlier result for $\mathbf{A}$ $(m\times n)$, $\mathbf{B}$ $(n\times p)$, and $\mathbf{C}$ $(p\times q)$:

$$\mathrm{tr}\{\mathbf{ABC}\} = (\mathrm{vec}(\mathbf{A}'))'(\mathbf{I}_m\otimes\mathbf{B})\,\mathrm{vec}(\mathbf{C}).$$

Making the relevant substitutions we find that

$$\begin{aligned} \mathbf{d}\,\mathrm{tr}\{\mathbf{AX}\} &= \mathbf{d}\,\mathrm{tr}\{\mathbf{A}\mathbf{I}_n\mathbf{X}\} = \mathbf{d}(\mathrm{vec}(\mathbf{A}'))'(\mathbf{I}_m\otimes\mathbf{I}_n)\,\mathrm{vec}(\mathbf{X}) \\ &= (\mathrm{vec}(\mathbf{A}'))'(\mathbf{I}_m\otimes\mathbf{I}_n)\,\mathbf{d}\,\mathrm{vec}(\mathbf{X}) = (\mathrm{vec}(\mathbf{A}'))'\,\mathbf{d}\,\mathrm{vec}(\mathbf{X}). \end{aligned}$$

Therefore, $\mathbf{J} = (\mathrm{vec}(\mathbf{A}'))'$.

Example 9.5.3. Differential of a Determinant.
In this example, we reconsider the differential of a determinant. Specifically, for $\mathbf{X}$ $m\times m$ and non-singular, and from the Laplace expansion for a determinant

$$\mathbf{d}\det(\mathbf{X}) = \sum_{j=1}^{m}\sum_{k=1}^{m} c_{jk}\,\mathrm{d}\,x_{jk} = (\mathrm{vec}(\boldsymbol{\Psi}))'\,\mathbf{d}\,\mathrm{vec}(\mathbf{X}) = \det(\mathbf{X})(\mathrm{vec}(\mathbf{X}))'\,\mathbf{d}\,\mathrm{vec}(\mathbf{X}),$$

where c_{jk} denotes the cofactor of x_{jk} and $\boldsymbol{\Psi}$ denotes the adjoint matrix of $\mathbf{X}$. Thus, $\mathbf{J} = \det(\mathbf{X})(\mathrm{vec}(\mathbf{X}))'$. The differenial of a scalar-valued function of matrix argument is one case where we have an alternate notation:

$$\frac{\partial\det(\mathbf{X})}{\partial\mathbf{X}} = \det(\mathbf{X})(\mathbf{X}')^{-1}.$$

Example 9.5.4. Differential of the Matrix Inverse.
Suppose that $\mathbf{X}$ is a non-singular $m\times m$ matrix and that $\mathbf{F}(\mathbf{X}) = \mathbf{X}^{-1}$. Observe that $\mathbf{X}^{-1}\mathbf{X} = \mathbf{I}_m$ and so, by the product rule,

$$\mathbf{0} = \mathbf{d}\,\mathbf{I}_m = \mathbf{d}(\mathbf{X}^{-1})\mathbf{X} + \mathbf{X}^{-1}\,\mathbf{d}\,\mathbf{X}$$

[6] Note the transpose in the denominator of the definition of $\mathbf{J}$ that was not their in our earlier result.

Table 9.1: Identification Table

Function	Differential	Jacobian Matrix $\mathbf{J}$	Order of $\mathbf{J}$
$\phi(\xi)$	$\mathbf{d}\,\phi = \alpha\,\mathbf{d}\,\xi$	α	1×1
$\phi(\mathbf{x})$	$\mathbf{d}\,\phi = \mathbf{a}'\,\mathbf{d}\,\mathbf{x}$	$\mathbf{a}'$	$1 \times n$
$\phi(\mathbf{X})$	$\mathbf{d}\,\phi = \mathrm{tr}\{\mathbf{A}'\,\mathbf{d}\,\mathbf{X}\} = (\mathrm{vec}(\mathbf{A}))'\,\mathbf{d}\,\mathrm{vec}(\mathbf{X})$	$(\mathrm{vec}(\mathbf{A}))'$	$1 \times nq$
$\mathbf{f}(\xi)$	$\mathbf{d}\,\mathbf{f} = \mathbf{a}\,\mathbf{d}\,\xi$	$\mathbf{a}$	$m \times 1$
$\mathbf{f}(\mathbf{x})$	$\mathbf{d}\,\mathbf{f} = \mathbf{A}\,\mathbf{d}\,\mathbf{x}$	$\mathbf{A}$	$m \times n$
$\mathbf{f}(\mathbf{X})$	$\mathbf{d}\,\mathbf{f} = \mathbf{A}\,\mathbf{d}\,\mathrm{vec}(\mathbf{x})$	$\mathbf{A}$	$n \times nq$
$\mathbf{F}(\xi)$	$\mathbf{d}\,\mathbf{F} = \mathbf{A}\,\mathbf{d}\,\xi$	$\mathrm{vec}(\mathbf{A})$	$mp \times 1$
$\mathbf{F}(\mathbf{x})$	$\mathbf{d}\,\mathrm{vec}(\mathbf{F}) = \mathbf{A}\,\mathbf{d}\,\mathbf{x}$	$\mathbf{A}$	$mp \times n$
$\mathbf{F}(\mathbf{X})$	$\mathbf{d}\,\mathrm{vec}(\mathbf{F}) = \mathbf{A}\,\mathbf{d}\,\mathrm{vec}(\mathbf{X})$	$\mathbf{A}$	$mp \times mp$

Notes: This is the first identification table from Magnus and Neudecker (1988). In this table, ϕ is a scalar-valued function, $\mathbf{f}$ an $m \times 1$ vector-valued function and $\mathbf{F}$ an $m \times p$ matrix function, while ξ is a scalar, $\mathbf{x}$ is an $n \times 1$ vector and $\mathbf{X}$ an $n \times q$ matrix. Finally, α is a scalar, $\mathbf{a}$ is a column vector and $\mathbf{A}$ is a matrix, each of which may be a function of $\mathbf{X}$, $\mathbf{x}$ or ξ.

$$\begin{aligned} &\Longrightarrow \mathbf{d}(\mathbf{X}^{-1})\mathbf{X} = -\mathbf{X}^{-1}\,\mathbf{d}\,\mathbf{X} \\ &\Longrightarrow \mathbf{d}(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(\mathbf{d}\,\mathbf{X})\mathbf{X}^{-1}, \\ &\Longrightarrow \mathrm{vec}(\mathbf{d}(\mathbf{X}^{-1})) = -\,\mathrm{vec}(\mathbf{X}^{-1}(\mathbf{d}\,\mathbf{X})\mathbf{X}^{-1}) = -((\mathbf{X}^{-1})' \otimes \mathbf{X}^{-1})\,\mathrm{vec}(\mathbf{d}\,\mathbf{X}). \end{aligned}$$

We conclude that $\mathbf{J} = -((\mathbf{X}^{-1})' \otimes \mathbf{X}^{-1})$. This is a very much simpler derivation than would be required if one tackled the problem directly (which would involve writing $\mathbf{X}^{-1}$ in terms of the ratio of its adjoint matrix $\mathbf{\Psi}$ to its determinant.

The ideas of some of these results can be found in Table 9.1. The way to read this table is to look for the type of function that you are working with, along with the type of argument that the function has. Then we see that when the relevant differentials can be written as linear functions of one another what the relevant Jacobian matrix is and what its dimensions will be. One special feature of the table to be aware of relates to the definition of the relationships between the differentials. These relate to the final forms of the relationships and so the various Jacobian matrices may be functions of the relevant arguments of the functions. The result for $\mathbf{d}\,\mathrm{tr}\{\mathbf{X}'\mathbf{X}\}$ is an example where the Jacobian matrix depends on $\mathbf{X}$.

9.6 Jacobians of Transformation

Jacobians of transformation are, perhaps, best thought of as scale factors required to maintain the appropriate size of the volume element when one makes changes of variables during intergration.[7] Specifically, changes of variables can change axes relative positions and can stretch and distort surfaces in the process. The role of the Jacobian of transformation is to adjust the volume element to take that into into account. That is twice that I have mentioned the volume element now. What exactly is it? If we write an arbitrary integral of some function $f(x_1, \ldots, x_m)$ over

[7] Jacobians of transformation are named for the German mathematician Carl Gustav Jacob Jacobi (1804–1851).

some region $A \in \mathbb{R}^m$ as

$$\int \cdots \int_A f(x_1, \ldots, x_m) \prod_{j=1}^{m} \mathrm{d}\, x_j,$$

then the volume element is the term

$$\prod_{j=1}^{m} \mathrm{d}\, x_j.$$

Now suppose that we make a set of m transformations from

$$\{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_m\}$$

with inverse transformations

$$\begin{aligned} x_1 &= x_1(y_1, \ldots, y_m) \\ &\vdots \\ x_m &= x_m(y_1, \ldots, y_m) \end{aligned} \qquad \text{or} \qquad \mathbf{x} = \mathbf{x}(\mathbf{y}).$$

That is, we are dealing with a situation where we have a vector-valued function of a vector, where both vectors are of the same dimension, so that the Jacobian matrix $\mathbf{J}$ is square. We shall restrict attention to the case where $\mathbf{J}$ is also non-singular, so that the various functions have partial derivatives that are linearly independenyt of one another. Then the Jacobian of transformation, denoted $J_{\mathbf{x}\to\mathbf{y}}$ is defined to be the absolute value of the determinant of the Jacobian matrix. That is,

$$J_{\mathbf{x}\to\mathbf{y}} = \operatorname{abs}\left(\det(\mathbf{J})\right), \qquad \text{where } \mathbf{J} = \frac{\partial \mathbf{x}'}{\partial \mathbf{y}}.$$

Given this definition we have our general change of variable formula:[8]

$$\int \cdots \int_A f(\mathbf{x}) \prod_{j=1}^{m} \mathrm{d}\, x_j = \int \cdots \int_{A'} f(\mathbf{x}(\mathbf{y}))\, J_{\mathbf{x}\to\mathbf{y}} \prod_{j=1}^{m} \mathrm{d}\, y_j,$$

where $A' \in \mathbb{R}^m$ denotes the image of A in the transformed space of $y_1, \ldots, y_m$. In this new coordinate system the volume element has become $J_{\mathbf{x}\to\mathbf{y}} \prod_{j=1}^{m} \mathrm{d}\, y_j$. We see from our earlier discussion that we can either find $J_{\mathbf{x}\to\mathbf{y}}$ directly, by calculating the relevant Jacobian matrix of partial derivatives, or by finding the matrix that defines the linear relationship between the differentials $\mathbf{d}\,\mathbf{x}$ and $\mathbf{d}\,\mathbf{y}$.

Example 9.6.1. Jacobians and the Bivariate Normal Distribution. Suppose that $\mathbf{Z} = [Z_1, Z_2]' \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_2)$ with joint density function

$$(2\pi)^{-1} \exp\{-\tfrac{1}{2}(Z_1^2 + Z_2^2)\} = (2\pi)^{-1} \exp\{-\tfrac{1}{2}\mathbf{Z}'\mathbf{Z}\}.$$

[8]This result is sometimes stated without the intergration, along the lines of

$$f(\mathbf{x}) = f(\mathbf{x}(\mathbf{y}))\, J_{\mathbf{x}\to\mathbf{y}}.$$

Now suppose that we define new variables $\mathbf{X} = [X_1, X_2]'$, with $X_1 = Z_1 + Z_2$ and $X_2 = Z_1 - Z_2$. We know that linear cobinations of Normal random variables are themseves Normal. Indeed, we see that

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{Z} \sim \mathrm{N}\left(\mathbf{0}, 2\mathbf{I}_2\right)$$

with corresponding joint density function

$$(2\pi)^{-1}(\det(2\mathbf{I}_2))^{-1/2} \exp\left\{-\tfrac{1}{2}\left(\tfrac{1}{2}\mathbf{X}'\mathbf{X}\right)\right\} = (4\pi)^{-1} \exp\left\{-\tfrac{1}{4}\mathbf{X}'\mathbf{X}\right\}. \tag{9.6.1}$$

Let us now try to obtain this result by direct substitution of our values for $\mathbf{X}$ into our original density function, i.e., our change of variable formula *without* the Jacobian of transformation. Noting that

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \mathbf{X} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{X} = \frac{1}{2} \begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix}. \tag{9.6.2}$$

Substituting this expression for $\mathbf{Z}$ into its joint density function yields

$$\begin{aligned} f(Z_1(\mathbf{X}), Z_2(\mathbf{X})) &= (2\pi)^{-1} \exp\left\{-\tfrac{1}{2} \times \tfrac{1}{4}\left((X_1 + X_2)^2 + (X_1 - X_2)^2\right)\right\} \\ &= (2\pi)^{-1} \exp\left\{-\tfrac{1}{4}\left(X_1^2 + X_2\right)\right\} \end{aligned}$$

which by comparison with (9.6.1) we recognise to be incorrect, as there is a factor of $\frac{1}{2}$ missing. That is, this expression would integrate to 2 rather than unity as required of a valid density function. So we need a scale factor to adjust the volume element appropriately. From the second member of (9.6.2), relation the differentials of $\mathbf{Z}$ and $\mathbf{x}$, we see immediately that the Jacobian of transformation is

$$\mathbf{J}_{Z\to\mathbf{X}} = \mathrm{abs}\left(\det(\mathbf{A})\right), \qquad \mathbf{A} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

That is, $\mathbf{J}_{Z\to\mathbf{X}} = \frac{1}{2}$, as required.

Note that in more complicated problems the Jacobian of transformation is likely to be a function of the set of variables that we are transforming into.

As a final remark, early significant contributions in this area of study were made by Deemer and Olkin (1951) and Olkin (1953).

9.7 Jacobians of Transformation and Exterior Differential Forms

The material of this section is quite a step-up from the earlier material and is presented for the sake of completenes. It was originally written as a free-standing document and so there may be some repetition of earlier material, not that there is anything wrong with that. The treatment given here is essentially that of Muirhead (1982, Chapter 2).

We have seen the rule telling us how to find the density of $\mathbf{y} = \mathbf{y}(\mathbf{x})$, $g(\mathbf{y})$ say, from the density of $\mathbf{x}$, $f(\mathbf{x})$ say, where $\mathbf{x}$ and $\mathbf{y}$ are both m-vectors; namely,

$$g(\mathbf{y}) = f(\mathbf{x}(\mathbf{y}))\,\mathrm{abs}\left(\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}'}\right)\right)$$

where $\mathbf{x} = \mathbf{x}(\mathbf{y})$ denotes the set of inverse functions relating the elements of $\mathbf{y}$ to the elements of $\mathbf{x}$, and

$$\frac{\partial \mathbf{x}'}{\partial \mathbf{y}} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \cdots & \frac{\partial x_m}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_m}{\partial y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial y_m} & \frac{\partial x_2}{\partial y_m} & \cdots & \frac{\partial x_m}{\partial y_m} \end{bmatrix}$$

denotes the Jacobian matrix. In high dimensional settings, $\partial \mathbf{x}'/\partial \mathbf{y}$ can be tedious to calculate and so we seek an alternative approach that will lead us to the same place but which may be a little less tedious to implement. As we shall see, the approach is based on differentials. Differentials have the advantage of being linear relationships and so can be relatively easy to work with.

By way of example, let's explore the particular case where $m = 2$ and suppose that our interest lies in evaluating the double integral

$$I = \iint_A f(x_1, x_2)\, \mathrm{d}\, x_1\, \mathrm{d}\, x_2, \tag{9.7.1}$$

where $A \subset \mathbb{R}^2$ denotes the region of integration. Let us further suppose that, in order to make progress, we need to make a change of variables from $\{x_1, x_2\} \to \{y_1, y_2\}$, for which the inverse relationships are

$$\begin{aligned} x_1 &= x_1(y_1, y_2) \\ x_2 &= x_2(y_1, y_2). \end{aligned}$$

After the change of variables we have

$$\begin{aligned} I &= \iint_{A'} f(x_1(y_1, y_2), x_2(y_1, y_2))\, \mathrm{abs}\left(\det \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{bmatrix}\right) \mathrm{d}\, y_1\, \mathrm{d}\, y_2 \\ &= \iint_{A'} f(x_1(y_1, y_2), x_2(y_1, y_2))\, \mathrm{abs}\left(\frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2}\frac{\partial x_2}{\partial y_1}\right) \mathrm{d}\, y_1\, \mathrm{d}\, y_2, \end{aligned} \tag{9.7.2}$$

where A' denotes the image of A after the transformation. So we see that we end up with the product of differential elements, $\mathrm{d}\, y_1$ and $\mathrm{d}\, y_2$, together with the Jacobian of transformation.

When $m = 2$ it is relatively simple to find the determinant of the Jacobian matrix but that will clearly become more of an issue as m increases. The use of differentials can help simplify this procedure in those cases but they do require some setup. Recall that the differential of the function $x_j = x_j(y_1, \ldots, y_m)$ is

$$\mathrm{d}\, x_j = \frac{\partial x_j}{\partial y_1}\, \mathrm{d}\, y_1 + \cdots + \frac{\partial x_j}{\partial y_m}\, \mathrm{d}\, y_m, \qquad j = 1, \ldots, m.$$

In the particular case of $m = 2$ we have

$$\begin{aligned} \mathrm{d}\, x_1 &= \frac{\partial x_1}{\partial y_1}\, \mathrm{d}\, y_1 + \frac{\partial x_1}{\partial y_2}\, \mathrm{d}\, y_2 \\ \mathrm{d}\, x_2 &= \frac{\partial x_2}{\partial y_1}\, \mathrm{d}\, y_1 + \frac{\partial x_2}{\partial y_2}\, \mathrm{d}\, y_2. \end{aligned}$$

Substituting these expressions into (9.7.1), when making the change of variables, yields the following expression for I

$$I = \iint_{A'} f(x_1(y_1, y_2), x_2(y_1, y_2)) \\ \times \left(\frac{\partial x_1}{\partial y_1} \, \mathrm{d}\, y_1 + \frac{\partial x_1}{\partial y_2} \, \mathrm{d}\, y_2 \right) \left(\frac{\partial x_2}{\partial y_1} \, \mathrm{d}\, y_1 + \frac{\partial x_2}{\partial y_2} \, \mathrm{d}\, y_2 \right). \quad (9.7.3)$$

In order for (9.7.2) and (9.7.3) to be the same it must be the case that, ignoring the absolute value function for a minute,

$$\left(\frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2} \frac{\partial x_1}{\partial y_m} \right) \mathrm{d}\, y_1 \, \mathrm{d}\, y_2 \\ = \left(\frac{\partial x_1}{\partial y_1} \, \mathrm{d}\, y_1 + \frac{\partial x_1}{\partial y_2} \, \mathrm{d}\, y_2 \right) \left(\frac{\partial x_2}{\partial y_1} \, \mathrm{d}\, y_1 + \frac{\partial x_2}{\partial y_2} \, \mathrm{d}\, y_2 \right). \quad (9.7.4)$$

To see what is required more clearly, expand the right-hand side to obtain

$$\frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_1} \, \mathrm{d}\, y_1 \, \mathrm{d}\, y_1 + \frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} \, \mathrm{d}\, y_1 \, \mathrm{d}\, y_2 + \frac{\partial x_1}{\partial y_2} \frac{\partial x_2}{\partial y_1} \, \mathrm{d}\, y_2 \, \mathrm{d}\, y_1 + \frac{\partial x_1}{\partial y_2} \frac{\partial x_2}{\partial y_2} \, \mathrm{d}\, y_2 \, \mathrm{d}\, y_2$$

and then compare this with the left-hand side of (9.7.4). What we observe is that the terms containing squares, such as $\mathrm{d}\, y_1 \, \mathrm{d}\, y_1$ and $\mathrm{d}\, y_2 \, \mathrm{d}\, y_2$, shouldn't be there and that cross-product terms of the form $\mathrm{d}\, y_1 \, \mathrm{d}\, y_2$ and $\mathrm{d}\, y_2 \, \mathrm{d}\, y_1$ should alternate in sign. Rather than simply using a regular product of the differentials, as illustrated above, in order to achieve this outcome it is necessary to use an anti-commutative or skew-symmetric product when multiplying the differentials together. Such a product satisfies these rules exactly, namely

$$\mathrm{d}\, y_1 \, \mathrm{d}\, y_2 = -\, \mathrm{d}\, y_2 \, \mathrm{d}\, y_1 \quad \text{and} \quad \mathrm{d}\, y_1 \, \mathrm{d}\, y_1 = \mathrm{d}\, y_2 \, \mathrm{d}\, y_2 = 0.$$

More generally, we can define this type of product according to

$$\mathrm{d}\, y_i \, \mathrm{d}\, y_j = -\, \mathrm{d}\, y_j \, \mathrm{d}\, y_i$$

where we note that if $i = j$ then the only value satisfying $\mathrm{d}\, y_j \, \mathrm{d}\, y_j = -\, \mathrm{d}\, y_j \, \mathrm{d}\, y_j$ is zero. In order to distinguish such a product from the more common commutative product, whereby $\mathrm{d}\, y_1 \, \mathrm{d}\, y_2 = \mathrm{d}\, y_2 \, \mathrm{d}\, y_1$, we write

$$\mathrm{d}\, y_i \wedge \mathrm{d}\, y_j = -\, \mathrm{d}\, y_j \wedge \mathrm{d}\, y_i.$$

These skew-symmetric products are formally called exterior products although, in recognition of the notation they are often referred to as a wedge product. Using this type of product we can write the right-hand side of (9.7.4) as

$$\left(\frac{\partial x_1}{\partial y_1} \frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2} \frac{\partial x_1}{\partial y_m} \right) \mathrm{d}\, y_1 \wedge \mathrm{d}\, y_2.$$

In this way the use of differentials has allowed us to recover the determinant that we need to properly evaluate the integral after the change of variables. The power of this observation lies in the following theorem.

Theorem. *If* $\mathrm{d}\,\mathbf{y}$ *is an* $m \times 1$ *vector of differentials and if* $\mathrm{d}\,\mathbf{x} = \mathbf{B}\,\mathrm{d}\,\mathbf{y}$, *where* $\mathbf{B}$ *is a an* $m \times m$ *non-singular matrix, then*

$$\bigwedge_{j=1}^{m} \mathrm{d}\,x_j = \det(\mathbf{B}) \bigwedge_{j=1}^{m} \mathrm{d}\,y_j.$$

The import of the theorem is that the relationship between $\mathrm{d}\,\mathbf{x}$ and $\mathrm{d}\,\mathbf{y}$ is a linear one and this may be easier to work with.

Chapter 10

Input-Output Matrices

One of the more important applications of matrix algebra in economics is that of input-output matrices. Indeed, Wassily W. Leontief was awarded the 1973 Nobel prize in economic science for the development of the 'input-output' method and its applications to economic problems. The problem addressed by input-output analysis is the following. Suppose that some of a producer's own product is consumed during the production process. What level of output must the producer produce in order to meet a given final demand, where final demand should be thought of as demand arising outside the production process for that good? For example, given that some electricity is consumed in the generation of electricity, how much electricity needs to be generated to meet a given final demand? This idea is illustrated in Figure 10.1. In this diagram, inputs enter the production process to become

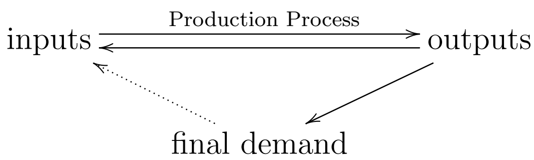


Figure 10.1: The Production Process

outputs as represented by the horizontal arrow moving from left to right. Some of the output produced goes to meet final demand but some of it may be used as inputs in the production process as indicated by the horizontal arrow from right to left. The dotted arrow is merely a recognition that some final demand may stem from other producers wanting to use the product as an input to their production process.

An essential feature of the problem described in the previous paragraph is that final demand is treated as a given, market forces play no role in the analysis.In a sense, it addresses questions of more interest in directed economies, such as some eastern European countries prior to the collapse of the communist bloc, or to a plant manager who has been told to achieve a given level of output for final sale. In what follows we shall make the following assumptions:

(i) Technical relationships of the economy remain fixed, i.e., inputs are consumed in the same proportion regardless of the level of production.

(ii) All output is consumed and output equals input. This does not preclude financial considerations because you can think of saving as a form of output and profit as a form of input.

(iii) Units of measurement are dollar values, so that everything is measured in the same units.

Because the technical relationships of the economy are assumed to remain fixed, various characteristics of the production process can be deduced from an input-output table. A major application of input-output tables in Australia is to deduce various parameter values for certain models of the economy — models which have been used for a variety of purposes including wage cases, determination of tariff rates, etc.

Consider Table 10.1 which provides an input-output table for a two industry economy. There are two sides to an input-output table, one being the producers

Table 10.1: Input-Output Table

Producers	Consumers (Input)			
(Output)	Industry A	Industry B	Final Demand	Totals
Industry A	40	120	40	200
Industry B	120	90	90	300
Other Factors	40	90		
Totals	200	300		

and the other being the consumers of output. Reading down the first column of Table 10.1 we see that Industry A and Industry B are both producers of output. Reading across the rows for these two producers tells us where their output goes. So, for example, Of the 200 units produced by Industry A, 40 are consumed by Industry A itself in its own production process, 120 units are consumed as inputs to the production process of Industry B and the the remaining 40 units of output go towards meeting Final Demand of consumers (which can be thought of as inputs to the production process of the consumers). The Other Factors (of production) is not a producer but incorporates such things as labour, profits and any other costs associated with the production process. Looking to the remaining columns, they tell us what is consumed in the production process. For example, Industry A consumes 40 units of output from Industry A, 120 units of output from Industry B and 40 units of other factors in its production process for a total of 200 units of factor inputs. (Remember the units of measurement are dollar values by assumption.) Similarly, Industry B consumes 300 units of factor inputs, consisting of 120 units from Industry A, 90 units from Industry B and 90 units of other factor inputs. The table is completed by row and column sums which tell us the total levels of production for each industry and the total consumption of inputs in the production process. By assumptions these totals must be equal for each industry.

Now that we know what the numbers are we can consider the problem that input-output tables are used to address. The Final Demand for each good is treated as a given and the problem is to work out the values that should appear in the Totals column (and row) in order to satisfy this level of demand. The remaining values in the body of the table are then immediately determined because of the fixed technical relationships in the production process between the levels of output and the required levels of inputs. That is, the proportions in which inputs need to be combined remain fixed, by assumption, at all levels of output.

Definition 10.1.1. Input-output coefficients.
Input-output coefficients are the proportions of total inputs consumed in a given industry that are spent on a given input. $\square$

Example 10.1.1. Calculating Input-Output Coefficients.
Input-output coefficients can be derived by dividing the elements of the top-left corner of an input-output table by the appropriate column sums. The input-output coefficients in Table 10.1 are:

$$\begin{bmatrix} \frac{40}{200} & \frac{120}{300} \\ \frac{120}{200} & \frac{90}{300} \\ \frac{40}{200} & \frac{90}{300} \end{bmatrix} = \begin{bmatrix} \frac{1}{5} & \frac{2}{5} \\ \frac{3}{5} & \frac{3}{10} \\ \frac{1}{5} & \frac{3}{10} \end{bmatrix}.$$

Note that, because the elements in a given column are proportions of total expenditure on the various inputs by that industry, the columns must sum to unity. $\square$

Definition 10.1.2. Input-Output Coefficient Matrix.
The *input-output coefficient matrix* is a square matrix containing those input-output coefficients relating to the various industries. That is, it gives the proportions of total inputs sourced from a given industry. $\square$

Example 10.1.2. Input-Output Coefficient Matrix.
In essence the input-output coefficient matrix is the matrix from the previous example excluding the row of input-output coefficients for the other factors of production. That is, the input-output coefficient matrix for Table 10.1 is

$$\begin{bmatrix} 0.2 & 0.4 \\ 0.6 & 0.3 \end{bmatrix}.$$

Thus, Industry A obtains 20% of its inputs from itself and 60% of its inputs from Industry B. Similarly, Industry B obtains 30% of its inputs from itself and 40% of them from Industry A. $\square$

The input-output coefficient matrix contains the parameter values that describe the technical relationships that make up the production process. From the original matrix of input-output data we see that the relationship encapsulated in the rows for Industry A and Industry B is

$$\text{value of output} = \text{value of inputs} + \text{value of final demand}.$$

That is, if $\mathbf{X}_A$ and $\mathbf{X}_B$ denote the total outputs of industries A and B, respectively, and if $\mathbf{D}_A$ and $\mathbf{D}_B$ denote the respective final demands for products of these industries, then

$$\mathbf{X}_A = 0.2\mathbf{X}_A + 0.4\mathbf{X}_B + \mathbf{D}_A \tag{10.1.1a}$$

$$\mathbf{X}_B = 0.6\mathbf{X}_A + 0.3\mathbf{X}_B + \mathbf{D}_B, \tag{10.1.1b}$$

or, in matrix notation,[1]

$$\mathbf{X} = \mathbf{F}\mathbf{X} + \mathbf{D}, \tag{10.1.2}$$

where $\mathbf{F}$ is the input-output coefficient matrix from before, $\mathbf{X}$ is the vector of total outputs and $\mathbf{D}$ is the vector of final demands. That is, from Equation (10.1.1a), Industry A is going to have to produce enough output so that it can meet its final demand ($\mathbf{D}_A$), satisfy the demand arising from Industry B (which will be 40% of the dollar value of whatever level of production Industry B chooses to operate at, something beyond the control of Industry A), and also meet its own needs which will be 20% of its final level of production. Similarly, from equation (10.1.1b), Industry B needs to produce enough output to meet its own needs (30% of its final level of output), its final demand and the demand arising from Industry A, which is 60% if the final level of output of Industry A. Rearranging Equation (10.1.2) yields

$$(\mathbf{I} - \mathbf{F})\,\mathbf{X} = \mathbf{D}.$$

$(\mathbf{I} - \mathbf{F})$ is called the *Leontief matrix.*

If $\mathbf{D} = \mathbf{0}$ then the model is said to be *closed.* A closed model is where no output leaves the industries, it is all used as inputs for further production. The system of equations is homogeneous and so it follows that $\mathbf{X} = \mathbf{0}$ unless $(\mathbf{I} - \mathbf{F})$ is singular, in which case an infinite number of solutions is possible.

If $\mathbf{D} \neq \mathbf{0}$ then the system is called an *open model* as some output leaves the system. That is, some output is consumed outside the industries. If a unique solution exists then it is given by $\mathbf{X} = (\mathbf{I} - \mathbf{F})^{-1}\,\mathbf{D}$.

Example 10.1.3. Determining Output Given Final Demand.
Using the input-output coefficients from Example 10.1.2, determine the levels of output in Industry A and Industry B required to meet final demands of 200 and 300, respectively.

SOLUTION

$$\mathbf{X} = (\mathbf{I} - \mathbf{F})^{-1}\,\mathbf{D} = \begin{bmatrix} 0.8 & -0.4 \\ -0.6 & 0.7 \end{bmatrix}^{-1} \begin{bmatrix} 200 \\ 300 \end{bmatrix} = \begin{bmatrix} 812.5 \\ 1125 \end{bmatrix}.$$

That is, $\mathbf{X}_A = 812.5$ units (some dollar value) and $\mathbf{X}_B = 1125$ units. Note that Industry A would also require $0.2\mathbf{X}_A = 812.5/5 = 162.5$ units of other factors of production, while Industry B requires 337.5 units of other factor inputs. □

[1] Observe that Equation (10.1.2) is of the form $\mathbf{\Gamma}\mathbf{X} = \mathbf{B}$, which has been discussed extensively throughout the earlier parts of this Chapter.

Bibliography

Abadir, K. M. and J. R. Magnus (2005). *Matrix Algebra, Econometric Exercises*, volume 1. Cambridge University Press. 104

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts. 104

Apostol, T. M. (1969). *Calculus: Volume 2*. John Wiley & SOns, New York, NY, second edition. 104

Courant, R. and F. John (1974). *Introduction to Calculus and Analysis: Volume II*, volume 2. John Wiley and Sons, Inc., New York. 104

Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*. Oxford University Press, New York. 69

Deemer, W. L. and I. Olkin (1951). The Jacobians of certain matrix transformations useful in multivariate analysis: Based on lectures of P. L. Hsu at the University of North Carolina, 1947. *Biometrika* 38(3/4), 345–367. 116

Funkhouser, H. G. (1930). A short account of the history of symmetric functions of roots of equations. *The American Mathematical Monthly* 37(7), 357–365, ISSN 00029890, 19300972. 87

Gupta, A. K. and D. K. Nagar (2000). *Matrix Variate Distributions*. Chapman & Hall/CRC, Boca Raton. 104

Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices. *SIAM Review* 23(1), 53–60, ISSN 00361445. 58, 61

Landau, E. (1965). *Differential and Integral Calculus*. AMS Chelsea Publishing, third edition, ISBN 978–0–8218–2830–4. American Mathematical Society Printing 2010. 87

Legendre, A.-M. (1805). *Nouvelles Méthodes Pour La Détermination Des Orbites Des Comètes*. F. Didot, Paris. Supplements were added in 1806 and 1820. 75

Legendre, A.-M. (1806). *Nouvelles Méthodes Pour La Détermination Des Orbites Des Comètes; Avec Un Supplément Contenant Divers Perfectionnemens De Ces Méthodes Et Leur Aux Deux Comètes De 1805*. Louis Courcier, Charles Crapelet, Paris. 75

Lütkepohl, H. (1996). *Handbook of Matrices*. John Wiley & Sons, Chichester, UK, ISBN 0471966886. 97, 104, 111

Maddala, G. S. (1992). *Introduction to Econometrics*. Macmillan Publishing Company, New York, second edition. 104

Magnus, J. R. (2017). *Introduction to the Theory of Econometrics*. VU University Press, Amsterdam. 104

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons Ltd, Chichester. 104, 112, 114

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, London. 104

Moskowitz, M. and F. Paliogiannis (2010). *Functions of Several Real Variables*. World Scientific, New Jersy. 104

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley and Sons, Inc., New York. 104, 116

Olkin, I. (1953). Note on 'The Jacobians of certain matrix transformations useful in multivariate analysis'. *Biometrika* 40(1/2), 43–46. 116

Ouellette, D. V. (1981). Schur complements and statistics. *Linear Algebra and Its Applications* 36, 187——295. 61, 93

Plackett, R. L. (1972). The discovery of the method of least squares. *Biometrika* 59(2), 239–251. 76

Pollock, D. S. G. (1979). *The Algebra of Econometrics*. John Wiley & Sons, Chichester, UK. 104

Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*. Oxford University Press, New York. 104

Searle, S. R. (1982). *Matrix Algebra Useful For Statistics*. John Wiley & Sons, New York. v, 78, 86, 100

Stigler, S. M. (1981). Gauss and the invention of least squares. *Annals of Statistics* 9(3), 465–474. 76

Theil, H. (1971). *Principles of Econometrics*. John Wiley and Sons, New York. 104