Appendix A

Matrix Algebra

A.1 Notation

A **scalar** a is a single number.

A **vector** $\boldsymbol{a}$ is a $k \times 1$ list of numbers, typically arranged in a column. We write this as

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

Equivalently, a vector $\boldsymbol{a}$ is an element of Euclidean k space, written as $\boldsymbol{a} \in \mathbb{R}^k$. If $k = 1$ then $\boldsymbol{a}$ is a scalar.

A **matrix** $\boldsymbol{A}$ is a $k \times r$ rectangular array of numbers, written as

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix}$$

By convention a_{ij} refers to the element in the $i'th$ row and $j'th$ column of $\boldsymbol{A}$. If $r = 1$ then $\boldsymbol{A}$ is a column vector. If $k = 1$ then $\boldsymbol{A}$ is a row vector. If $r = k = 1$, then $\boldsymbol{A}$ is a scalar.

A standard convention (which we will follow in this text whenever possible) is to denote scalars by lower-case italics (a), vectors by lower-case bold italics ($\boldsymbol{a}$), and matrices by upper-case bold italics ($\boldsymbol{A}$). Sometimes a matrix $\boldsymbol{A}$ is denoted by the symbol (a_{ij}).

A matrix can be written as a set of column vectors or as a set of row vectors. That is,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_r \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{bmatrix}$$

where

$$\boldsymbol{a}_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ki} \end{bmatrix}$$

are column vectors and

$$\boldsymbol{\alpha}_j = \begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jr} \end{bmatrix}$$

are row vectors.

The **transpose** of a matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}'$, $\boldsymbol{A}^\top$, or $\boldsymbol{A}^t$, is obtained by flipping the matrix on its diagonal. Thus

$$\boldsymbol{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

Alternatively, letting $\boldsymbol{B} = \boldsymbol{A}'$, then $b_{ij} = a_{ji}$. Note that if $\boldsymbol{A}$ is $k \times r$, then $\boldsymbol{A}'$ is $r \times k$. If $\boldsymbol{a}$ is a $k \times 1$ vector, then $\boldsymbol{a}'$ is a $1 \times k$ row vector.

A matrix is **square** if $k = r$. A square matrix is **symmetric** if $\boldsymbol{A} = \boldsymbol{A}'$, which requires $a_{ij} = a_{ji}$. A square matrix is **diagonal** if the off-diagonal elements are all zero, so that $a_{ij} = 0$ if $i \neq j$. A square matrix is **upper (lower) diagonal** if all elements below (above) the diagonal equal zero.

An important diagonal matrix is the **identity matrix**, which has ones on the diagonal. The $k \times k$ identity matrix is denoted as

$$\boldsymbol{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

A **partitioned matrix** takes the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1r} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2r} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{A}_{k1} & \boldsymbol{A}_{k2} & \cdots & \boldsymbol{A}_{kr} \end{bmatrix}$$

where the A_{ij} denote matrices, vectors and/or scalars.

A.2 Complex Matrices*

Scalars, vectors and matrices may contain real or complex numbers as entries. (However, most econometric applications exclusively use real matrices.) If all elements of a vector $\boldsymbol{x}$ are real we say that $\boldsymbol{x}$ is a real vector, and similarly for matrices.

Recall that a complex number can be written as $x = a + b\mathrm{i}$ where where $\mathrm{i} = \sqrt{-1}$ and a and b are real numbers. Similarly a vector with complex elements can be written as

$$\boldsymbol{x} = \boldsymbol{a} + \boldsymbol{b}\mathrm{i}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are real vectors, and a matrix with complex elements can be written as

$$\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{B}\mathrm{i}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are real matrices.

Recall that the complex conjugate of $x = a+b\mathrm{i}$ is $x^* = a-b\mathrm{i}$. For matrices, the analgous concept is the conjugate transpose. The conjugate transpose of $\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{B}\mathrm{i}$ is $\boldsymbol{X}^* = \boldsymbol{A}' - \boldsymbol{B}'\mathrm{i}$. It is obtained by taking the transpose and taking the complex conjugate of each element.

A.3 Matrix Addition

If the matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ are of the same order, we define the sum

$$\boldsymbol{A} + \boldsymbol{B} = (a_{ij} + b_{ij}) .$$

Matrix addition follows the commutative and associative laws:

$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$$
$$\boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}) = (\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C}.$$

A.4 Matrix Multiplication

If $\boldsymbol{A}$ is $k \times r$ and c is real, we define their product as

$$\boldsymbol{A}c = c\boldsymbol{A} = (a_{ij}c) .$$

If $\boldsymbol{a}$ and $\boldsymbol{b}$ are both $k \times 1$, then their inner product is

$$\boldsymbol{a}'\boldsymbol{b} = a_1b_1 + a_2b_2 + \cdots + a_kb_k = \sum_{j=1}^{k} a_jb_j.$$

Note that $\boldsymbol{a}'\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{a}$. We say that two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are **orthogonal** if $\boldsymbol{a}'\boldsymbol{b} = 0$.

If $\boldsymbol{A}$ is $k \times r$ and $\boldsymbol{B}$ is $r \times s$, so that the number of columns of $\boldsymbol{A}$ equals the number of rows of $\boldsymbol{B}$, we say that $\boldsymbol{A}$ and $\boldsymbol{B}$ are **conformable**. In this event the matrix product $\boldsymbol{AB}$ is defined. Writing $\boldsymbol{A}$ as a set of row vectors and $\boldsymbol{B}$ as a set of column vectors (each of length r), then the matrix product

is defined as

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \\ \vdots \\ \boldsymbol{a}_k' \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_s \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_s \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_s \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_s \end{bmatrix}.$$

Matrix multiplication is not commutative: in general $\boldsymbol{AB} \neq \boldsymbol{BA}$**.** However, it is associative and distributive:

$$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$$
$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$$

An alternative way to write the matrix product is to use matrix partitions. For example,

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

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

As another example,

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_r \end{bmatrix}$$
$$= \boldsymbol{A}_1\boldsymbol{B}_1 + \boldsymbol{A}_2\boldsymbol{B}_2 + \cdots + \boldsymbol{A}_r\boldsymbol{B}_r$$
$$= \sum_{j=1}^{r} \boldsymbol{A}_j\boldsymbol{B}_j$$

An important property of the identity matrix is that if $\boldsymbol{A}$ is $k \times r$, then $\boldsymbol{A}\boldsymbol{I}_r = \boldsymbol{A}$ and $\boldsymbol{I}_k\boldsymbol{A} = \boldsymbol{A}$.

The $k \times r$ matrix $\boldsymbol{A}$, $r \leq k$, is called **orthonormal** if $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_r$.

A.5 Trace

The **trace** of a $k \times k$ square matrix $\boldsymbol{A}$ is the sum of its diagonal elements

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} a_{ii}.$$

Some straightforward properties for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and real c are

$$\begin{aligned}
\operatorname{tr}(c\boldsymbol{A}) &= c\operatorname{tr}(\boldsymbol{A}) \\
\operatorname{tr}\left(\boldsymbol{A}'\right) &= \operatorname{tr}(\boldsymbol{A}) \\
\operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) &= \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}) \\
\operatorname{tr}(\boldsymbol{I}_k) &= k.
\end{aligned}$$

Also, for $k \times r$ $\boldsymbol{A}$ and $r \times k$ $\boldsymbol{B}$ we have

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}). \tag{A.1}$$

Indeed,

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) &= \operatorname{tr}\begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_k \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_k \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_k \end{bmatrix} \\
&= \sum_{i=1}^{k} \boldsymbol{a}_i'\boldsymbol{b}_i \\
&= \sum_{i=1}^{k} \boldsymbol{b}_i'\boldsymbol{a}_i \\
&= \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}).
\end{aligned}$$

A.6 Rank and Inverse

The rank of the $k \times r$ matrix $(r \leq k)$

$$\boldsymbol{A} = \left[\begin{array}{cccc} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_r \end{array} \right]$$

is the number of linearly independent columns $\boldsymbol{a}_j$, and is written as $\text{rank}\,(\boldsymbol{A})$. We say that $\boldsymbol{A}$ has full rank if $\text{rank}\,(\boldsymbol{A}) = r$.

A square $k \times k$ matrix $\boldsymbol{A}$ is said to be **nonsingular** if it is has full rank, e.g. $\text{rank}\,(\boldsymbol{A}) = k$. This means that there is no $k \times 1$ $\boldsymbol{c} \neq \boldsymbol{0}$ such that $\boldsymbol{Ac} = \boldsymbol{0}$.

If a square $k \times k$ matrix $\boldsymbol{A}$ is nonsingular then there exists a unique matrix $k \times k$ matrix $\boldsymbol{A}^{-1}$ called the **inverse** of $\boldsymbol{A}$ which satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k.$$

For non-singular $\boldsymbol{A}$ and $\boldsymbol{C}$, some important properties include

$$\begin{aligned} \boldsymbol{A}\boldsymbol{A}^{-1} &= \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k \\ \left(\boldsymbol{A}^{-1}\right)' &= \left(\boldsymbol{A}'\right)^{-1} \\ (\boldsymbol{AC})^{-1} &= \boldsymbol{C}^{-1}\boldsymbol{A}^{-1} \\ (\boldsymbol{A}+\boldsymbol{C})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{C}^{-1} \\ \boldsymbol{A}^{-1} - (\boldsymbol{A}+\boldsymbol{C})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{A}^{-1} \end{aligned}$$

Also, if $\boldsymbol{A}$ is an orthonormal matrix, then $\boldsymbol{A}^{-1} = \boldsymbol{A}'$.

Another useful result for non-singular $\boldsymbol{A}$ is known as the **Woodbury matrix identity**

$$(\boldsymbol{A}+\boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{C}\left(\boldsymbol{C}+\boldsymbol{CDA}^{-1}\boldsymbol{BC}\right)^{-1}\boldsymbol{CDA}^{-1}. \quad \text{(A.2)}$$

In particular, for $\boldsymbol{C} = -1$, $\boldsymbol{B} = \boldsymbol{b}$ and $\boldsymbol{D} = \boldsymbol{b}'$ for vector $\boldsymbol{b}$ we find what is known as the **Sherman–Morrison formula**

$$\left(\boldsymbol{A}-\boldsymbol{bb}'\right)^{-1} = \boldsymbol{A}^{-1} + \left(1-\boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1}\boldsymbol{A}^{-1}\boldsymbol{bb}'\boldsymbol{A}^{-1}. \quad \text{(A.3)}$$

The following fact about inverting partitioned matrices is quite useful.

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}^{-1} \stackrel{def}{=} \begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11\cdot 2}^{-1} & -\boldsymbol{A}_{11\cdot 2}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ -\boldsymbol{A}_{22\cdot 1}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{A}_{22\cdot 1}^{-1} \end{bmatrix} \tag{A.4}$$

where $\boldsymbol{A}_{11\cdot 2} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$ and $\boldsymbol{A}_{22\cdot 1} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$. There are alternative algebraic representations for the components. For example, using the Woodbury matrix identity you can show the following alternative expressions

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

Even if a matrix $\boldsymbol{A}$ does not possess an inverse, we can still define the **Moore-Penrose generalized inverse** $\boldsymbol{A}^{-}$ as the matrix which satisfies

$$\begin{aligned} \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A} &= \boldsymbol{A} \\ \boldsymbol{A}^{-}\boldsymbol{A}\boldsymbol{A}^{-} &= \boldsymbol{A}^{-} \\ \boldsymbol{A}\boldsymbol{A}^{-} &\text{ is symmetric} \\ \boldsymbol{A}^{-}\boldsymbol{A} &\text{ is symmetric} \end{aligned}$$

For any matrix $\boldsymbol{A}$, the Moore-Penrose generalized inverse $\boldsymbol{A}^{-}$ exists and is unique.

For example, if

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

and when $\boldsymbol{A}_{11}^{-1}$ exists then

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

A.7 Determinant

The **determinant** is a measure of the volume of a square matrix. It is written as $\det \boldsymbol{A}$ or $|\boldsymbol{A}|$.

While the determinant is widely used, its precise definition is rarely needed. However, we present the definition here for completeness. Let $\boldsymbol{A} = (a_{ij})$ be a $k \times k$ matrix . Let $\pi = (j_1, ..., j_k)$ denote a permutation of $(1, ..., k)$. There are $k!$ such permutations. There is a unique count of the number of inversions of the indices of such permutations (relative to the natural order $(1, ..., k)$, and let $\varepsilon_\pi = +1$ if this count is even and $\varepsilon_\pi = -1$ if the count is odd. Then the determinant of $\boldsymbol{A}$ is defined as

$$\det \boldsymbol{A} = \sum_{\pi} \varepsilon_\pi a_{1j_1} a_{2j_2} \cdots a_{kj_k}.$$

For example, if $\boldsymbol{A}$ is 2×2, then the two permutations of $(1, 2)$ are $(1, 2)$ and $(2, 1)$, for which $\varepsilon_{(1,2)} = 1$ and $\varepsilon_{(2,1)} = -1$. Thus

$$\begin{aligned} \det \boldsymbol{A} &= \varepsilon_{(1,2)} a_{11} a_{22} + \varepsilon_{(2,1)} a_{21} a_{12} \\ &= a_{11} a_{22} - a_{12} a_{21}. \end{aligned}$$

For a square matrix $\boldsymbol{A}$, the **minor** M_{ij} of the ij^{th} element a_{ij} is the determinant of the matrix obtained by removing the i^{th} row and j^{th} column of $\boldsymbol{A}$. The **cofactor** of the ij^{th} element is $C_{ij} = (-1)^{i+j} M_{ij}$. An important representation known as Laplace's expansion, relates the determinant of $\boldsymbol{A}$ to its cofactors:

$$\det \boldsymbol{A} = \sum_{j=1}^{k} a_{ij} C_{ij}$$

This holds for all $i = 1, , .., k$. This is often presented as a method for computation of a determinant.

Some properties of the determinant include

- $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$
- $\det(c\boldsymbol{A}) = c^k \det \boldsymbol{A}$
- $\det(\boldsymbol{AB}) = (\det \boldsymbol{A})(\det \boldsymbol{B})$

- $\det\left(\boldsymbol{A}^{-1}\right) = (\det \boldsymbol{A})^{-1}$
- $\det\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = (\det \boldsymbol{D}) \det\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)$ if $\det \boldsymbol{D} \neq 0$
- $\det \boldsymbol{A} \neq 0$ if and only if $\boldsymbol{A}$ is nonsingular
- If $\boldsymbol{A}$ is triangular (upper or lower), then $\det \boldsymbol{A} = \prod_{i=1}^{k} a_{ii}$
- If $\boldsymbol{A}$ is orthogonal, then $\det \boldsymbol{A} = \pm 1$
- $\boldsymbol{A}^{-1} = (\det \boldsymbol{A})^{-1} \boldsymbol{C}$ where $\boldsymbol{C} = (C_{ij})$ is the matrix of cofactors

A.8 Eigenvalues

The characteristic equation of a $k \times k$ square matrix $\boldsymbol{A}$ is

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}_k) = 0.$$

The left side is a polynomial of degree k in λ so it has exactly k roots, which are not necessarily distinct and may be real or complex. They are called the **latent roots** or **characteristic roots** or **eigenvalues** of $\boldsymbol{A}$. If λ_i is an eigenvalue of $\boldsymbol{A}$, then $\boldsymbol{A} - \lambda_i \boldsymbol{I}_k$ is singular so there exists a non-zero vector $\boldsymbol{h}_i$ such that

$$(\boldsymbol{A} - \lambda_i \boldsymbol{I}_k)\, \boldsymbol{h}_i = \boldsymbol{0}.$$

The vector $\boldsymbol{h}_i$ is called a **latent vector** or **characteristic vector** or **eigenvector** of $\boldsymbol{A}$ corresponding to λ_i.

We now state some useful properties. Let λ_i and $\boldsymbol{h}_i$, $i = 1, ..., k$ denote the k eigenvalues and eigenvectors of a square matrix $\boldsymbol{A}$. Let $\boldsymbol{\Lambda}$ be a diagonal matrix with the characteristic roots in the diagonal, and let $\boldsymbol{H} = [\boldsymbol{h}_1 \cdots \boldsymbol{h}_k]$.

- $\det(\boldsymbol{A}) = \prod_{i=1}^{k} \lambda_i$
- $\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} \lambda_i$
- $\boldsymbol{A}$ is non-singular if and only if all its characteristic roots are non-zero.
- If $\boldsymbol{A}$ has distinct characteristic roots, there exists a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{A} = \boldsymbol{P}^{-1}\boldsymbol{\Lambda}\boldsymbol{P}$ and $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1} = \boldsymbol{\Lambda}$.

- If $k \times k$ $\boldsymbol{A}$ is symmetric, then $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ and $\boldsymbol{H}'\boldsymbol{A}\boldsymbol{H} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues on the diagaonal, and $\boldsymbol{H}'\boldsymbol{A}\boldsymbol{H} = \boldsymbol{I}_k$. The characteristic roots are all real. $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ is called the **spectral decomposition** of $\boldsymbol{A}$.

- When the eigenvalues of $k \times k$ $\boldsymbol{A}$ are real it is conventional to write them in decending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. We also write $\lambda_{\min}(\boldsymbol{A}) = \lambda_k = \min\{\lambda_\ell\}$ and $\lambda_{\max}(\boldsymbol{A}) = \lambda_1 = \max\{\lambda_\ell\}$.

- For real symmetric $\boldsymbol{A}$, $\lambda_{\max}(\boldsymbol{A}) = \max_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$

- For real symmetric $\boldsymbol{A}$, $\lambda_{\min}(\boldsymbol{A}) = \min_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$

- The characteristic roots of $\boldsymbol{A}^{-1}$ are λ_1^{-1}, λ_2^{-1}, ..., λ_k^{-1}.

- The matrix $\boldsymbol{H}$ has the **orthonormal** properties $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}$.

- $\boldsymbol{H}^{-1} = \boldsymbol{H}'$ and $(\boldsymbol{H}')^{-1} = \boldsymbol{H}$

- For any $k \times 1$ vector $\boldsymbol{a}$, $\lambda_{\max}(\boldsymbol{a}\boldsymbol{a}') = \boldsymbol{a}'\boldsymbol{a}$

A.9 Positive Definiteness

We say that a $k \times k$ real symmetric square matrix $\boldsymbol{A}$ is **positive semi-definite** if for all $\boldsymbol{c} \neq \boldsymbol{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \geq 0$. This is written as $\boldsymbol{A} \geq 0$. We say that $\boldsymbol{A}$ is **positive definite** if for all $\boldsymbol{c} \neq \boldsymbol{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} > 0$. This is written as $\boldsymbol{A} > 0$.

Some properties include:

- If $\boldsymbol{A} = \boldsymbol{G}'\boldsymbol{B}\boldsymbol{G}$ with $\boldsymbol{B} \geq 0$ and some matrix $\boldsymbol{G}$**,** then $\boldsymbol{A}$ is positive semi-definite. (For any $\boldsymbol{c} \neq \boldsymbol{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} = \boldsymbol{\alpha}'\boldsymbol{B}\boldsymbol{\alpha} \geq 0$ where $\boldsymbol{\alpha} = \boldsymbol{G}\boldsymbol{c}$.) If $\boldsymbol{G}$ has full column rank and $\boldsymbol{B} > 0$, then $\boldsymbol{A}$ is positive definite.

- If $\boldsymbol{A}$ is positive definite, then $\boldsymbol{A}$ is non-singular and $\boldsymbol{A}^{-1}$ exists. Furthermore, $\boldsymbol{A}^{-1} > 0$.

- $\boldsymbol{A} > 0$ if and only if it is symmetric and all its characteristic roots are positive.

- By the spectral decomposition, $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is diagonal with non-negative diagonal elements. All diagonal elements of $\boldsymbol{\Lambda}$ are strictly positive if (and only if) $\boldsymbol{A} > 0$.
- The rank of $\boldsymbol{A}$ equals the number of strictly positive characteristic roots.
- If $\boldsymbol{A} > 0$ then $\boldsymbol{A}^{-1} = \boldsymbol{H}\boldsymbol{\Lambda}^{-1}\boldsymbol{H}'$.
- If $\boldsymbol{A} \geq 0$ and $\operatorname{rank}(\boldsymbol{A}) = r < k$ then $\boldsymbol{A}^{-} = \boldsymbol{H}\boldsymbol{\Lambda}^{-}\boldsymbol{H}'$ where $\boldsymbol{A}^{-}$ is the Moore-Penrose generalized inverse, and $\boldsymbol{\Lambda}^{-} = \operatorname{diag}\left(\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_k^{-1}, 0, ..., 0\right)$
- If $\boldsymbol{A} \geq 0$ we can find a matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}'$. We call $\boldsymbol{B}$ a **matrix square root** of $\boldsymbol{A}$. The matrix $\boldsymbol{B}$ need not be unique. One way to construct $\boldsymbol{B}$ is to use the spectral decomposition $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda}$ is diagonal, and then set $\boldsymbol{B} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}$. There is a unique square root $\boldsymbol{B}$ which is also positive semi-definite $\boldsymbol{B} \geq 0$. If $\boldsymbol{A} > 0$ then $\boldsymbol{B} > 0$.

A $k \times k$ square matrix $\boldsymbol{A}$ is **idempotent** if $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. If $\boldsymbol{A}$ is idempotent and symmetric with rank r, then it has r characteristic roots which equal 1 and $k - r$ characteristic roots which equal 0. To see this, by the spectral decomposition that we can write $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}$ is orthogonal and $\boldsymbol{\Lambda}$ contains the eigenvalues. Then

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

We deduce that $\boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}$ and $\lambda_i^2 = \lambda_i$ for $i = 1, ..., k$. Hence the λ_i must equal either 0 or 1. Since the rank of $\boldsymbol{A}$ is r, and the rank equals the number of positive characteristic roots, it follows that

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}_{k-r} \end{bmatrix}$$

and the spectral decomposition of idempotent $\boldsymbol{A}$ takes the form

$$\boldsymbol{A} = \boldsymbol{H} \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}_{k-r} \end{bmatrix} \boldsymbol{H}' \tag{A.5}$$

with $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$. Additionally, $\operatorname{tr}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A})$ and $\boldsymbol{A}$ is positive semi-definite.

If $\boldsymbol{A}$ is **idempotent** then $\boldsymbol{I}-\boldsymbol{A}$ is also idempotent.

One useful fact is that if $\boldsymbol{A}$ is idempotent then for any conformable vector $\boldsymbol{c}$,

$$\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \leq \boldsymbol{c}'\boldsymbol{c} \tag{A.6}$$

$$\boldsymbol{c}'\left(\boldsymbol{I}-\boldsymbol{A}\right)\boldsymbol{c} \leq \boldsymbol{c}'\boldsymbol{c} \tag{A.7}$$

To see this, note that

$$\boldsymbol{c}'\boldsymbol{c} = \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} + \boldsymbol{c}'\left(\boldsymbol{I}-\boldsymbol{A}\right)\boldsymbol{c}.$$

Since $\boldsymbol{A}$ and $\boldsymbol{I}-\boldsymbol{A}$ are idempotent, they are both positive semi-definite, so both $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}$ and $\boldsymbol{c}'\left(\boldsymbol{I}-\boldsymbol{A}\right)\boldsymbol{c}$ are non-negative. Thus they must satisfy (A.6)-(A.7).

A.10 Singular Values

The singular values of a $k \times r$ real matrix $\boldsymbol{A}$ are the square roots of the eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$. Thus for $j = 1, ..., r$

$$s_j = \sqrt{\lambda_j\left(\boldsymbol{A}'\boldsymbol{A}\right)}$$

Since $\boldsymbol{A}'\boldsymbol{A}$ is positive semi-definite, its eigenvalues are non-negative so the singular values are real and non-negative.

The non-zero singular values of $\boldsymbol{A}$ and $\boldsymbol{A}'$ are the same.

When $\boldsymbol{A}$ is positive semi-definite then the singular values of $\boldsymbol{A}$ correspond to its eigenvalues.

The singular value decomposition of a $k \times r$ real matrix $\boldsymbol{A}$ takes the form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}'$ where $\boldsymbol{U}$ is $k \times k$, $\boldsymbol{\Lambda}$ is $k \times r$ and $\boldsymbol{V}$ is $r \times r$, with $\boldsymbol{U}$ and $\boldsymbol{V}$ orthonormal ($\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_k$ and $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_r$) and $\boldsymbol{\Lambda}$ is a diagonal matrix with the singular values of $\boldsymbol{A}$ on the diagonal.

It is convention to write the singular values in decending order $s_1 \geq s_2 \geq \cdots \geq s_r$.

A.11 Matrix Calculus

Let $\boldsymbol{x} = (x_1, ..., x_k)$ be $k \times 1$ and $g(\boldsymbol{x}) = g(x_1, ..., x_k) : \mathbb{R}^k \to \mathbb{R}$. The vector derivative is

$$\frac{\partial}{\partial \boldsymbol{x}} g(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(\boldsymbol{x}) \\ \vdots \\ \frac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{pmatrix}$$

and

$$\frac{\partial}{\partial \boldsymbol{x}'} g(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(\boldsymbol{x}) & \cdots & \frac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{pmatrix}.$$

Some properties are now summarized.

- $\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{a}'\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{x}'\boldsymbol{a}) = \boldsymbol{a}$
- $\frac{\partial}{\partial \boldsymbol{x}'} (\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}$
- $\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}$
- $\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}'} (\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A} + \boldsymbol{A}'$
- $\frac{\partial}{\partial \boldsymbol{A}} \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}) = \boldsymbol{B}'$
- $\frac{\partial}{\partial \boldsymbol{A}} \log\det(\boldsymbol{A}) = \left(\boldsymbol{A}^{-}\right)'$

The final two results require some justification. Recall from Section A.5 that we can write out explicitly

$$\operatorname{tr}(\boldsymbol{B}\boldsymbol{A}) = \sum_i \sum_j a_{ij} b_{ji}.$$

Thus if we take the derivative with respect to a_{ij} we find

$$\frac{\partial}{\partial a_{ij}} \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}) = b_{ji}.$$

which is the ij^{th} element of $\boldsymbol{B}'$, establishing the second-to-last result.

For the final result, recall Laplace's expansion

$$\det \boldsymbol{A} = \sum_{j=1}^{k} a_{ij} C_{ij}$$

where C_{ij} is the ij^{th} cofactor of $\boldsymbol{A}$. Set $\boldsymbol{C} = (C_{ij})$. Observe that C_{ij} for $j = 1, ..., k$ are not functions of a_{ij}. Thus the derivative with respect to a_{ij} is

$$\frac{\partial}{\partial a_{ij}} \log \det (\boldsymbol{A}) = (\det \boldsymbol{A})^{-1} \frac{\partial}{\partial a_{ij}} \det \boldsymbol{A} = (\det \boldsymbol{A})^{-1} C_{ij}$$

Together this implies

$$\frac{\partial}{\partial \boldsymbol{A}} \log \det (\boldsymbol{A}) = (\det \boldsymbol{A})^{-1} \boldsymbol{C} = \boldsymbol{A}^{-1}$$

where the second equality is a property of the inverse from Section A.7.

A.12 Kronecker Products and the Vec Operator

Let $\boldsymbol{A} = [\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ \cdots \ \boldsymbol{a}_n]$ be $m \times n$. The **vec** of $\boldsymbol{A}$, denoted by $\text{vec}(\boldsymbol{A})$, is the $mn \times 1$ vector

$$\text{vec}(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix}.$$

Let $\boldsymbol{A} = (a_{ij})$ be an $m \times n$ matrix and let $\boldsymbol{B}$ be any matrix. The **Kronecker product** of $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted $\boldsymbol{A} \otimes \boldsymbol{B}$, is the matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

Some important properties are now summarized. These results hold for matrices for which all matrix multiplications are conformable.

- $(\boldsymbol{A} + \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{C} + \boldsymbol{B} \otimes \boldsymbol{C}$

- $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$
- $\boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes C$
- $(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'$
- $\operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \operatorname{tr}(\boldsymbol{B})$
- If $\boldsymbol{A}$ is $m \times m$ and $\boldsymbol{B}$ is $n \times n$, $\det(\boldsymbol{A} \otimes \boldsymbol{B}) = (\det(\boldsymbol{A}))^n (\det(\boldsymbol{B}))^m$
- $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$
- If $\boldsymbol{A} > 0$ and $\boldsymbol{B} > 0$ then $\boldsymbol{A} \otimes \boldsymbol{B} > 0$
- $\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B})$
- $\operatorname{tr}(\boldsymbol{ABCD}) = \operatorname{vec}(\boldsymbol{D}')' (\boldsymbol{C}' \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B})$

A.13 Vector Norms

Given any vector space V (such as Euclidean space $\mathbb{R}^m$) a **norm** on V is a function $\rho : V \to \mathbb{R}$ with the properties

1. $\rho(c\boldsymbol{a}) = |c| \rho(\boldsymbol{a})$ for any complex number c and $\boldsymbol{a} \in V$
2. $\rho(\boldsymbol{a} + \boldsymbol{b}) \leq \rho(\boldsymbol{a}) + \rho(\boldsymbol{b})$
3. If $\rho(\boldsymbol{a}) = 0$ then $\boldsymbol{a} = \boldsymbol{0}$

A seminorm on V is a function which satisfies the first two properties. The second property is known as the triangle inequality, and it is the one property which typically needs a careful demonstration (as the other two properties typically hold by inspection).

The typical norm used for Euclidean space $\mathbb{R}^m$ is the **Euclidean norm**

$$\begin{aligned} \|\boldsymbol{a}\| &= \left(\boldsymbol{a}'\boldsymbol{a}\right)^{1/2} \\ &= \left(\sum_{i=1}^{m} a_i^2\right)^{1/2}. \end{aligned}$$

Alternative norms include the $p-$norm (for $p \geq 1$)

$$\|\boldsymbol{a}\|_p = \left(\sum_{i=1}^{m} |a_i|^p \right)^{1/p}$$

Special cases include the Euclidean norm ($p = 2$), the $1-$norm

$$\|\boldsymbol{a}\|_1 = \sum_{i=1}^{m} |a_i|$$

and the sup-norm

$$\|\boldsymbol{a}\|_\infty = \max \left(|a_1|, ..., |a_m| \right).$$

For real numbers ($m = 1$) these norms coincide.

Some standard inequalities for Euclidean space are now given. The Minkowski inequality given below establishes that any p-norm with $p \geq 1$ (including the Euclidean norm) satisfies the triangle inequality and is thus a valid norm.

Jensen's Inequality. If $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is convex, then for any non-negative weights a_j such that $\sum_{j=1}^{m} a_j = 1$, and any real numbers x_j

$$g \left(\sum_{j=1}^{m} a_j x_j \right) \leq \sum_{j=1}^{m} a_j g \left(x_j \right). \tag{A.8}$$

In particular, setting $a_j = 1/m$, then

$$g \left(\frac{1}{m} \sum_{j=1}^{m} x_j \right) \leq \frac{1}{m} \sum_{j=1}^{m} g \left(x_j \right). \tag{A.9}$$

If $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is concave then the inequalities in (A.8) and (A.9) are reversed.

Weighted Geometric Mean Inequality. For any non-negative real weights a_j such that $\sum_{j=1}^{m} a_j = 1$, and any non-negative real numbers x_j

$$x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m} \leq \sum_{j=1}^{m} a_j x_j \tag{A.10}$$

Loève's c_r Inequality. For $r > 0$,

$$\left|\sum_{j=1}^{m} a_j\right|^r \leq c_r \sum_{j=1}^{m} |a_j|^r \tag{A.11}$$

where $c_r = 1$ when $r \leq 1$ and $c_r = m^{r-1}$ when $r \geq 1$.

c_2 Inequality. For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$(\boldsymbol{a}+\boldsymbol{b})'(\boldsymbol{a}+\boldsymbol{b}) \leq 2\boldsymbol{a}'\boldsymbol{a} + 2\boldsymbol{b}'\boldsymbol{b} \tag{A.12}$$

Hölder's Inequality. If $p > 1$, $q > 1$, and $1/p + 1/q = 1$, then for any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\sum_{j=1}^{m} |a_j b_j| \leq \|\boldsymbol{a}\|_p \|\boldsymbol{b}\|_q \tag{A.13}$$

Minkowski's Inequality. For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, if $p \geq 1$, then

$$\|\boldsymbol{a}+\boldsymbol{b}\|_p \leq \|\boldsymbol{a}\|_p + \|\boldsymbol{b}\|_p \tag{A.14}$$

Schwarz Inequality. For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\left|\boldsymbol{a}'\boldsymbol{b}\right| \leq \|\boldsymbol{a}\| \|\boldsymbol{b}\|. \tag{A.15}$$

Proof of Jensen's Inequality (A.8). By the definition of convexity, for any $\lambda \in [0,1]$

$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda) g(x_2). \tag{A.16}$$

This implies

$$\begin{aligned} g\left(\sum_{j=1}^{m} a_j x_j\right) &= g\left(a_1 x_1 + (1-a_1)\sum_{j=2}^{m} \frac{a_j}{1-a_1} x_j\right) \\ &\leq a_1 g(x_1) + (1-a_1)\, g\left(\sum_{j=2}^{m} b_j x_j\right). \end{aligned}$$

where $b_j = a_j/(1-a_1)$ and $\sum_{j=2}^m b_j = 1$. By another application of (A.16) this is bounded by

$$a_1 g(x_1) + (1-a_1)\left(b_2 g(x_2) + (1-b_2) g\left(\sum_{j=2}^m c_j x_j\right)\right)$$
$$= a_1 g(x_1) + a_2 g(x_2) + (1-a_1)(1-b_2) g\left(\sum_{j=2}^m c_j x_j\right)$$

where $c_j = b_j/(1-b_2)$. By repeated application of (A.16) we obtain (A.8). ■

Proof of Weighted Geometric Mean Inequality. Since the logarithm is strictly concave, by Jensen's inequality

$$\log\left(x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m}\right) = \sum_{j=1}^m a_j \log x_j \le \log\left(\sum_{j=1}^m a_j x_j\right).$$

Applying the exponential yields (A.10). ■

Proof of Loève's c_r Inequality. For $r \ge 1$ this is simply a rewriting of the finite form Jensen's inequality (A.9) with $g(u) = u^r$. For $r < 1$, define $b_j = |a_j| \,/\, \left(\sum_{j=1}^m |a_j|\right)$. The facts that $0 \le b_j \le 1$ and $r < 1$ imply $b_j \le b_j^r$ and thus

$$1 = \sum_{j=1}^m b_j \le \sum_{j=1}^m b_j^r$$

which implies

$$\left(\sum_{j=1}^m |a_j|\right)^r \le \sum_{j=1}^m |a_j|^r.$$

The proof is completed by observing that

$$\left(\sum_{j=1}^m a_j\right)^r \le \left(\sum_{j=1}^m |a_j|\right)^r.$$

■

Proof of c_2 Inequality. By the c_r inequality, $(a_j + b_j)^2 \leq 2a_j^2 + 2b_j^2$. Thus

$$\begin{aligned}(\boldsymbol{a}+\boldsymbol{b})'(\boldsymbol{a}+\boldsymbol{b}) &= \sum_{j=1}^{m}(a_j+b_j)^2 \\ &\leq 2\sum_{j=1}^{m}a_j^2 + 2\sum_{j=1}^{m}b_j^2 \\ &= 2\boldsymbol{a}'\boldsymbol{a} + 2\boldsymbol{b}'\boldsymbol{b}\end{aligned}$$

■

Proof of Hölder's Inequality. Set $u_j = |a_j|^p / \|\boldsymbol{a}\|_p^p$ and $u_j = |b_j|^q / \|\boldsymbol{b}\|_q^q$ and observe that $\sum_{j=1}^m u_j = 1$ and $\sum_{j=1}^m v_j = 1$. By the weighted geometric mean inequality,

$$u_j^{1/p} v_j^{1/q} \leq \frac{u_j}{p} + \frac{v_j}{q}.$$

Then since $\sum_{j=1}^m u_j = 1$, $\sum_{j=1}^m v_j = 1$ and $1/p + 1/q = 1$

$$\frac{\sum_{j=1}^m |a_j b_j|}{\|\boldsymbol{a}\|_p \|\boldsymbol{b}\|_q} = \sum_{j=1}^{m} u_j^{1/p} v_j^{1/q} \leq \sum_{j=1}^{m}\left(\frac{u_j}{p} + \frac{v_j}{q}\right) = 1$$

which is (A.13). ■

Proof of Minkowski's Inequality. Se $q = p/(p-1)$ so that $1/p + 1/q = 1$. Using the triangle inequality for real numbers and two applications of Hölder's

inequality

$$\begin{aligned}
\|\boldsymbol{a}+\boldsymbol{b}\|_p^p &= \sum_{j=1}^{m} |a_j+b_j|^p \\
&= \sum_{j=1}^{m} |a_j+b_j|\,|a_j+b_j|^{p-1} \\
&\le \sum_{j=1}^{m} |a_j|\,|a_j+b_j|^{p-1} + \sum_{j=1}^{m} |b_j|\,|a_j+b_j|^{p-1} \\
&\le \|\boldsymbol{a}\|_p \left(\sum_{j=1}^{m} |a_j+b_j|^{(p-1)q}\right)^{1/q} + \|\boldsymbol{b}\|_p \left(\sum_{j=1}^{m} |a_j+b_j|^{(p-1)q}\right)^{1/q} \\
&= \left(\|\boldsymbol{a}\|_p + \|\boldsymbol{b}\|_p\right) \|\boldsymbol{a}+\boldsymbol{b}\|_p^{p-1}
\end{aligned}$$

Solving, we find (A.14). ■

Proof of Schwarz Inequality. Using Hölder's inequality with $p = q = 2$

$$\left|\boldsymbol{a}'\boldsymbol{b}\right| \le \sum_{j=1}^{m} |a_j b_j| \le \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|$$

■

A.14 Matrix Norms

Two common norms used for matrix spaces are the **Frobenius norm** and the **spectral norm**. We can write either as $\|\boldsymbol{A}\|$, but may write $\|\boldsymbol{A}\|_F$ or $\|\boldsymbol{A}\|_2$ when we want to be specific.

The **Frobenius norm** of an $m \times k$ matrix $\boldsymbol{A}$ is the Euclidean norm applied

to its elements

$$\begin{aligned}\|\boldsymbol{A}\|_F &= \|\mathrm{vec}\,(\boldsymbol{A})\| \\ &= \left(\mathrm{tr}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} \\ &= \left(\sum_{i=1}^{m}\sum_{j=1}^{k} a_{ij}^2\right)^{1/2}.\end{aligned}$$

If an $m \times m$ real matrix $\boldsymbol{A}$ is symmetric with eigenvalues λ_ℓ, $\ell = 1, ..., m$, then

$$\|\boldsymbol{A}\|_F = \left(\sum_{\ell=1}^{m} \lambda_\ell^2\right)^{1/2}.$$

To see this, by the spectral decomposition $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ with $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, ..., \lambda_m\}$, so

$$\|\boldsymbol{A}\|_F = \left(\mathrm{tr}\left(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'\right)\right)^{1/2} = \left(\mathrm{tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}\right)\right)^{1/2} = \left(\sum_{\ell=1}^{m} \lambda_\ell^2\right)^{1/2}. \quad \text{(A.17)}$$

A useful calculation is for any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, using (A.1),

$$\left\|\boldsymbol{a}\boldsymbol{b}'\right\|_F = \mathrm{tr}\left(\boldsymbol{b}\boldsymbol{a}'\boldsymbol{a}\boldsymbol{b}'\right)^{1/2} = \left(\boldsymbol{b}'\boldsymbol{b}\boldsymbol{a}'\boldsymbol{a}\right)^{1/2} = \|\boldsymbol{a}\|\,\|\boldsymbol{b}\| \quad \text{(A.18)}$$

and in particular

$$\left\|\boldsymbol{a}\boldsymbol{a}'\right\|_F = \|\boldsymbol{a}\|^2. \quad \text{(A.19)}$$

The **spectral norm** of an $m \times k$ real matrix $\boldsymbol{A}$ is its largest singular value

$$\|\boldsymbol{A}\|_2 = s_{\max}(\boldsymbol{A}) = \left(\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2}$$

where $\lambda_{\max}(\boldsymbol{B})$ denotes the largest eigenvalue of the matrix $\boldsymbol{B}$. Notice that

$$\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right) = \left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2$$

so

$$\|\boldsymbol{A}\|_2 = \left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2^{1/2}.$$

If $\boldsymbol{A}$ is $m \times m$ and symmetric with eigenvalues λ_j then

$$\|\boldsymbol{A}\|_2 = \max_{j \leq m} |\lambda_j| .$$

The Frobenius and spectral norms are closely related. They are equivalent when applied to a matrix of rank 1, since $\left\|\boldsymbol{a}\boldsymbol{b}'\right\|_2 = \|\boldsymbol{a}\| \|\boldsymbol{b}\| = \left\|\boldsymbol{a}\boldsymbol{b}'\right\|_F$. In general, for $m \times k$ matrix $\boldsymbol{A}$ with rank r

$$\|\boldsymbol{A}\|_2 = \left(\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} \leq \left(\sum_{j=1}^{k} \lambda_j\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} = \|\boldsymbol{A}\|_F .$$

Since $\boldsymbol{A}'\boldsymbol{A}$ also has rank at most r, it has at most r non-zero eigenvalues, and hence

$$\|\boldsymbol{A}\|_F = \left(\sum_{j=1}^{k} \lambda_j\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} = \left(\sum_{j=1}^{r} \lambda_j\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} \leq \left(r\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} = \sqrt{r}\,\|\boldsymbol{A}\|_2$$

Given any vector norm $\|\boldsymbol{a}\|$ the **induced matrix norm** is defined as

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\| = \sup_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

To see that this is a norm we need to check that it satisfies the triangle inequality. Indeed

$$\|\boldsymbol{A}+\boldsymbol{B}\| = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{x}\| \leq \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\| + \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{B}\boldsymbol{x}\| = \|\boldsymbol{A}\| + \|\boldsymbol{B}\| .$$

For any vector $\boldsymbol{x}$, by the definition of the induced norm

$$\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\| \|\boldsymbol{x}\|$$

a property which is called consistent norms.

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be conformable and $\|\boldsymbol{A}\|$ an induced matrix norm. Then using the property of consistent norms

$$\|\boldsymbol{A}\boldsymbol{B}\| = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\| \leq \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\| \|\boldsymbol{B}\boldsymbol{x}\| = \|\boldsymbol{A}\| \|\boldsymbol{B}\| ,$$

A matrix norm which satisfies this property is called a **sub-multiplicative norm**, and is a matrix form of the Schwarz inequality.

Of particular interest, the matrix norm induced by the Euclidean vector norm is the spectral norm. Indeed,

$$\sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\|^2 = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{x} = \lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right) = \|\boldsymbol{A}\|_2^2 .$$

It follows that the spectral norm is consistent with the Euclidean norm, and is sub-multiplicative.

A.15 Matrix Inequalities

Schwarz Matrix Inequality: For any $m \times k$ and $k \times m$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and either the Frobenius or spectral norm,

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{B}\| . \tag{A.20}$$

Triangle Inequality: For any $m \times k$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and either the Frobenius or spectral norm,

$$\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\| . \tag{A.21}$$

Trace Inequality. For any $m \times m$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that $\boldsymbol{A}$ is symmetric and $\boldsymbol{B} \geq 0$

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \leq \|\boldsymbol{A}\|_2 \operatorname{tr}(\boldsymbol{B}) . \tag{A.22}$$

Quadratic Inequality. For any $m \times 1$ $\boldsymbol{b}$ and $m \times m$ symmetric matrix $\boldsymbol{A}$

$$\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b} \leq \|\boldsymbol{A}\|_2 \, \boldsymbol{b}'\boldsymbol{b} \tag{A.23}$$

Strong Schwarz Matrix Inequality. For any conformable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$

$$\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_2 \, \|\boldsymbol{B}\|_F . \tag{A.24}$$

Norm Equivalence. For any $m \times k$ matrix $\boldsymbol{A}$ of rank r

$$\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F \leq \sqrt{r} \, \|\boldsymbol{A}\|_2 . \tag{A.25}$$

Eigenvalue Product Inequality. For any $m \times m$ real symmetric matrices $\boldsymbol{A} \geq 0$ and $\boldsymbol{B} \geq 0$, the eigenvalues $\lambda_\ell(\boldsymbol{AB})$ are real and satisfy

$$\lambda_{\min}(\boldsymbol{A})\lambda_{\min}(\boldsymbol{B}) \leq \lambda_\ell(\boldsymbol{AB}) = \lambda_\ell\left(\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}\right) \leq \lambda_{\max}(\boldsymbol{A})\lambda_{\max}(\boldsymbol{B}) \tag{A.26}$$

(Zhang and Zhang, 2006, Corollary 11)

Proof of Schwarz Matrix Inequality: The inequality holds for the spectral norm since it is an induced norm. Now consider the Frobenius norm. Partition $\boldsymbol{A}' = [\boldsymbol{a}_1, ..., \boldsymbol{a}_n]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_n]$. Then by partitioned matrix multiplication, the definition of the Frobenius norm and the Schwarz inequality for vectors

$$\begin{aligned}
\|\boldsymbol{AB}\|_F &= \left\| \begin{array}{ccc} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots \\ \vdots & \vdots & \ddots \end{array} \right\|_F \\
&\leq \left\| \begin{array}{ccc} \|\boldsymbol{a}_1\|\|\boldsymbol{b}_1\| & \|\boldsymbol{a}_1\|\|\boldsymbol{b}_2\| & \cdots \\ \|\boldsymbol{a}_2\|\|\boldsymbol{b}_1\| & \|\boldsymbol{a}_2\|\|\boldsymbol{b}_2\| & \cdots \\ \vdots & \vdots & \ddots \end{array} \right\|_F \\
&= \left(\sum_{i=1}^m \sum_{j=1}^m \|\boldsymbol{a}_i\|^2 \|\boldsymbol{b}_j\|^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^m \|\boldsymbol{a}_i\|^2 \right)^{1/2} \left(\sum_{i=1}^m \|\boldsymbol{b}_i\|^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^k \sum_{j=1}^m \boldsymbol{a}_{ji}^2 \right)^{1/2} \left(\sum_{i=1}^m \sum_{j=1}^k \|\boldsymbol{b}_{ji}\|^2 \right)^{1/2} \\
&= \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F
\end{aligned}$$

■

Proof of Triangle Inequality: The inequality holds for the spectral norm since it is an induced norm. Now consider the Frobenius norm. Let $\boldsymbol{a} =$

$\text{vec}(\boldsymbol{A})$ and $\boldsymbol{b} = \text{vec}(\boldsymbol{B})$. Then by the definition of the Frobenius norm and the Schwarz Inequality for vectors

$$\begin{aligned}\|\boldsymbol{A}+\boldsymbol{B}\|_F &= \|\text{vec}(\boldsymbol{A}+\boldsymbol{B})\|_F \\ &= \|\boldsymbol{a}+\boldsymbol{b}\| \\ &\leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\| \\ &= \|\boldsymbol{A}\|_F + \|\boldsymbol{B}\|_F\end{aligned}$$

■

Proof of Trace Inequality. By the spectral decomposition for symmetric matices, $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda}$ has the eigenvalues λ_j of $\boldsymbol{A}$ on the diagonal and $\boldsymbol{H}$ is orthonormal. Define $\boldsymbol{C} = \boldsymbol{H}'\boldsymbol{B}\boldsymbol{H}$ which has non-negative diagonal elements C_{jj} since $\boldsymbol{B}$ is positive semi-definite. Then

$$\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{\Lambda}\boldsymbol{C}) = \sum_{j=1}^{m} \lambda_j C_{jj} \leq \max_j |\lambda_j| \sum_{j=1}^{m} C_{jj} = \|\boldsymbol{A}\|_2 \,\text{tr}(\boldsymbol{C})$$

where the inequality uses the fact that $C_{jj} \geq 0$. But note that

$$\text{tr}(\boldsymbol{C}) = \text{tr}\left(\boldsymbol{H}'\boldsymbol{B}\boldsymbol{H}\right) = \text{tr}\left(\boldsymbol{H}\boldsymbol{H}'\boldsymbol{B}\right) = \text{tr}(\boldsymbol{B})$$

since $\boldsymbol{H}$ is orthonormal. Thus $\text{tr}(\boldsymbol{A}\boldsymbol{B}) \leq \|\boldsymbol{A}\|_2 \,\text{tr}(\boldsymbol{B})$ as stated. ■

Proof of Quadratic Inequality: In the Trace Inequality set $\boldsymbol{B} = \boldsymbol{b}\boldsymbol{b}'$ and note $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \boldsymbol{b}'\boldsymbol{A}\boldsymbol{b}$ and $\text{tr}(\boldsymbol{B}) = \boldsymbol{b}'\boldsymbol{b}$. ■

Proof of Strong Schwarz Matrix Inequality. By the definition of the Frobenius norm, the property of the trace, the Trace Inequality (noting that both $\boldsymbol{A}'\boldsymbol{A}$ and $\boldsymbol{B}\boldsymbol{B}'$ are symmetric and positive semi-definite), and the Schwarz matrix inequality

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{B}\|_F &= \left(\text{tr}\left(\boldsymbol{B}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\right)\right)^{1/2} \\ &= \left(\text{tr}\left(\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}'\right)\right)^{1/2} \\ &\leq \left(\left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2 \text{tr}\left(\boldsymbol{B}\boldsymbol{B}'\right)\right)^{1/2} \\ &= \|\boldsymbol{A}\|_2 \,\|\boldsymbol{B}\|_F .\end{aligned}$$

■

Appendix B

Probability

B.1 Foundations

The set S of all possible outcomes of an experiment is called the **sample space** for the experiment. Take the simple example of tossing a coin. There are two outcomes, heads and tails, so we can write $S = \{H, T\}$. If two coins are tossed in sequence, we can write the four outcomes as $S = \{HH, HT, TH, TT\}$.

An **event** A is any collection of possible outcomes of an experiment. An event is a subset of S, including S itself and the null set $\emptyset$. Continuing the two coin example, one event is $A = \{HH, HT\}$, the event that the first coin is heads. We say that A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$. For example, the sets $\{HH, HT\}$ and $\{TH\}$ are disjoint. Furthermore, if the sets $A_1, A_2, ...$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, ...$ is called a **partition** of S.

The following are elementary set operations:

Union: $A \cup B = \{x : x \in A \text{ or } x \in B\}$.

Intersection: $A \cap B = \{x : x \in A \text{ and } x \in B\}$.

Complement: $A^c = \{x : x \notin A\}$.

The following are useful properties of set operations.

Commutatitivity: $A \cup B = B \cup A$; $\quad A \cap B = B \cap A$.

Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$; $\quad A \cap (B \cap C) = (A \cap B) \cap C$.

Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $\quad A \cup (B \cap C) =$

$(A \cup B) \cap (A \cup C)$.

DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$; $\quad (A \cap B)^c = A^c \cup B^c$.

A **probability function** assigns probabilities (numbers between 0 and 1) to events A in S. This is straightforward when S is countable; when S is uncountable we must be somewhat more careful. A set $\mathcal{B}$ is called a **sigma algebra** (or Borel field) if $\emptyset \in \mathcal{B}$, $A \in \mathcal{B}$ implies $A^c \in \mathcal{B}$, and $A_1, A_2, ... \in \mathcal{B}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$. A simple example is $\{\emptyset, S\}$ which is known as the trivial sigma algebra. For any sample space S, let $\mathcal{B}$ be the smallest sigma algebra which contains all of the open sets in S. When S is countable, $\mathcal{B}$ is simply the collection of all subsets of S, including $\emptyset$ and S. When S is the real line, then $\mathcal{B}$ is the collection of all open and closed intervals. We call $\mathcal{B}$ the sigma algebra associated with S. We only define probabilities for events contained in $\mathcal{B}$.

We now can give the axiomatic definition of probability. Given S and $\mathcal{B}$, a probability function Pr satisfies $\Pr(S) = 1$, $\Pr(A) \geq 0$ for all $A \in \mathcal{B}$, and if $A_1, A_2, ... \in \mathcal{B}$ are pairwise disjoint, then $\Pr\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$.

Some important properties of the probability function include the following

- $\Pr(\emptyset) = 0$
- $\Pr(A) \leq 1$
- $\Pr(A^c) = 1 - \Pr(A)$
- $\Pr(B \cap A^c) = \Pr(B) - \Pr(A \cap B)$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- If $A \subset B$ then $\Pr(A) \leq \Pr(B)$
- Bonferroni's Inequality: $\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$
- Boole's Inequality: $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$

For some elementary probability models, it is useful to have simple rules to count the number of objects in a set. These counting rules are facilitated by using the binomial coefficients which are defined for nonnegative integers n and r, $n \geq r$, as

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

When counting the number of objects in a set, there are two important distinctions. Counting may be **with replacement** or **without replacement**. Counting may be **ordered** or **unordered**. For example, consider a lottery where you pick six numbers from the set 1, 2, ..., 49. This selection is without replacement if you are not allowed to select the same number twice, and is with replacement if this is allowed. Counting is ordered or not depending on whether the sequential order of the numbers is relevant to winning the lottery. Depending on these two distinctions, we have four expressions for the number of objects (possible arrangements) of size r from n objects.

	Without Replacement	With Replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

In the lottery example, if counting is unordered and without replacement, the number of potential combinations is $\binom{49}{6} = 13,983,816$.

If $\Pr(B) > 0$ the **conditional probability** of the event A given the event B is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

For any B, the conditional probability function is a valid probability function where S has been replaced by B. Rearranging the definition, we can write

$$\Pr(A \cap B) = \Pr(A \mid B) \Pr(B)$$

which is often quite useful. We can say that the occurrence of B has no information about the likelihood of event A when $\Pr(A \mid B) = \Pr(A)$, in which case we find

$$\Pr(A \cap B) = \Pr(A) \Pr(B) \tag{B.1}$$

We say that the events A and B are **statistically independent** when (B.1) holds. Furthermore, we say that the collection of events $A_1, ..., A_k$ are **mutually independent** when for any subset $\{A_i : i \in I\}$,

$$\Pr\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \Pr(A_i).$$

Theorem 1 *(Bayes' Rule). For any set B and any partition $A_1, A_2, ...$ of the sample space, then for each $i = 1, 2, ...$*

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i)\Pr(A_i)}{\sum_{j=1}^{\infty}\Pr(B \mid A_j)\Pr(A_j)}$$

B.2 Random Variables

A **random variable** X is a function from a sample space S into the real line. This induces a new sample space – the real line – and a new probability function on the real line. Typically, we denote random variables by uppercase letters such as X, and use lower case letters such as x for potential values and realized values. (This is in contrast to the notation adopted for most of the textbook.) For a random variable X we define its **cumulative distribution function** (CDF) as

$$F(x) = \Pr(X \leq x). \tag{B.2}$$

Sometimes we write this as $F_X(x)$ to denote that it is the CDF of X. A function $F(x)$ is a CDF if and only if the following three properties hold:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$
2. $F(x)$ is nondecreasing in x
3. $F(x)$ is right-continuous

We say that the random variable X is **discrete** if $F(x)$ is a step function. In the latter case, the range of X consists of a countable set of real numbers $\tau_1, ..., \tau_r$. The probability function for X takes the form

$$\Pr(X = \tau_j) = \pi_j, \qquad j = 1, ..., r \tag{B.3}$$

where $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{r} \pi_j = 1$.

We say that the random variable X is **continuous** if $F(x)$ is continuous in x. In this case $\Pr(X = \tau) = 0$ for all $\tau \in R$ so the representation (B.3) is unavailable. Instead, we represent the relative probabilities by the **probability density function** (PDF)

$$f(x) = \frac{d}{dx}F(x)$$

so that

$$F(x) = \int_{-\infty}^{x} f(u)du$$

and

$$\Pr(a \leq X \leq b) = \int_{a}^{b} f(u)du.$$

These expressions only make sense if $F(x)$ is differentiable. While there are examples of continuous random variables which do not possess a PDF, these cases are unusual and are typically ignored.

A function $f(x)$ is a PDF if and only if $f(x) \geq 0$ for all $x \in R$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.

B.3 Expectation

For any real function g, we define the **mean** or **expectation** $\mathbb{E}g(X)$ as follows. If X is discrete,

$$\mathbb{E}(g(X)) = \sum_{j=1}^{r} g(\tau_j)\pi_j,$$

and if X is continuous

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

The latter is well defined and finite if

$$\int_{-\infty}^{\infty} |g(x)| f(x)dx < \infty. \tag{B.4}$$

If (B.4) does not hold, evaluate

$$I_1 = \int_{g(x)>0} g(x)f(x)dx$$

$$I_2 = -\int_{g(x)<0} g(x)f(x)dx$$

If $I_1 = \infty$ and $I_2 < \infty$ then we define $\mathbb{E}(g(X)) = \infty$. If $I_1 < \infty$ and $I_2 = \infty$ then we define $\mathbb{E}(g(X)) = -\infty$. If both $I_1 = \infty$ and $I_2 = \infty$ then $\mathbb{E}(g(X))$ is undefined.

Since $\mathbb{E}(a + bX) = a + b\mathbb{E}(X)$, we say that expectation is a linear operator.

For $m > 0$, we define the m^{th} **moment** of X as $\mathbb{E}(X^m)$ and the m^{th} **central moment** as $\mathbb{E}((X - \mathbb{E}X)^m)$.

Two special moments are the **mean** $\mu = \mathbb{E}(X)$ and **variance** $\sigma^2 = \mathbb{E}(X - \mu)^2 = \mathbb{E}(X^2) - \mu^2$. We call $\sigma = \sqrt{\sigma^2}$ the **standard deviation** of X. We can also write $\sigma^2 = \mathrm{var}(X)$. For example, this allows the convenient expression $\mathrm{var}(a + bX) = b^2 \mathrm{var}(X)$.

The **moment generating function** (MGF) of X is

$$M(\lambda) = \mathbb{E}(\exp(\lambda X)).$$

The MGF does not necessarily exist. However, when it does and $\mathbb{E}(|X|^m) < \infty$ then

$$\frac{d^m}{d\lambda^m} M(\lambda)\bigg|_{\lambda=0} = \mathbb{E}(X^m)$$

which is why it is called the moment generating function.

More generally, the **characteristic function** (CF) of X is

$$C(\lambda) = \mathbb{E}(\exp(\mathrm{i}\lambda X))$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. The CF always exists, and when $\mathbb{E}(|X|^m) < \infty$

$$\frac{d^m}{d\lambda^m} C(\lambda)\bigg|_{\lambda=0} = \mathrm{i}^m \mathbb{E}(X^m).$$

The L^p **norm**, $p \geq 1$, of the random variable X is

$$\|X\|_p = (\mathbb{E}(|X|^p))^{1/p}.$$

B.4 Gamma Function

The gamma function is defined for $\alpha > 0$ as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x)\, dx. \tag{B.5}$$

By integration by parts you can show that it satisfies the property

$$\Gamma(1+\alpha) = \Gamma(\alpha)\alpha.$$

Thus for positive integers n,

$$\Gamma(n) = (n-1)!$$

Hence the gamma function can be viewed as a continuous version of the factorial.

Special values include

$$\Gamma(1) = \int_0^\infty \exp(-x)\, dx = 1 \tag{B.6}$$

and

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{B.7}$$

The latter holds by making the change of variables $x = u^2$ in (B.5) and applying (5.2).

A useful fact is

$$\int_0^\infty y^{a-1} \exp(-by)\, dy = b^{-a}\Gamma(a) \tag{B.8}$$

which can be found by applying change-of-variables to the definition (B.5).

Another is for for $\alpha \in \mathbb{R}$

$$\lim_{n\to\infty} \frac{\Gamma(n+\alpha)}{\Gamma(n)\, n^\alpha} = 1. \tag{B.9}$$

Sterling's formula is an expansion for the the logarithm of the gamma function (and hence for the factorial as well).

$$\log\Gamma(\alpha) = \frac{1}{2}\log(2\pi) + \left(\alpha - \frac{1}{2}\right)\log\alpha - z + \frac{1}{12\alpha} - \frac{1}{360\alpha^3} + \frac{1}{1260\alpha^5} + \cdots$$

B.5 Common Distributions

For reference, we now list some important discrete distribution function.

Bernoulli

$$\begin{aligned}\Pr(X=x) &= p^x(1-p)^{1-x}, \qquad x=0,1; \qquad 0\le p\le 1\\ \mathbb{E}(X) &= p\\ \operatorname{var}(X) &= p(1-p)\end{aligned}$$

Binomial

$$\begin{aligned}\Pr(X=x) &= \binom{n}{x}p^x(1-p)^{n-x}, \qquad x=0,1,...,n; \qquad 0\le p\le 1\\ \mathbb{E}(X) &= np\\ \operatorname{var}(X) &= np(1-p)\end{aligned}$$

Geometric

$$\begin{aligned}\Pr(X=x) &= p(1-p)^{x-1}, \qquad x=1,2,...; \qquad 0\le p\le 1\\ \mathbb{E}(X) &= \frac{1}{p}\\ \operatorname{var}(X) &= \frac{1-p}{p^2}\end{aligned}$$

Multinomial

$$\begin{aligned}\Pr(X_1=x_1, X_2=x_2, ..., X_m=x_m) &= \frac{n!}{x_1!x_2!\cdots x_m!}p_1^{x_1}p_2^{x_2}\cdots p_m^{x_m},\\ x_1+\cdots+x_m &= n;\\ p_1+\cdots+p_m &= 1\\ \mathbb{E}(X_i) &= p_i\\ \operatorname{var}(X_i) &= np_i(1-p_i)\\ \operatorname{cov}(X_i,X_j) &= -np_ip_j\end{aligned}$$

Negative Binomial

$$\Pr(X=x)=\frac{\Gamma(r+x)}{x!\Gamma(r)}p^r(1-p)^{x-1}, \qquad x=0,1,2,...; \qquad 0\le p\le 1$$
$$\mathbb{E}(X)=\frac{r(1-p)}{p}$$
$$\mathrm{var}(X)=\frac{r(1-p)}{p^2}$$

Poisson

$$\Pr(X=x)=\frac{\exp(-\lambda)\lambda^x}{x!}, \qquad x=0,1,2,..., \qquad \lambda>0$$
$$\mathbb{E}(X)=\lambda$$
$$\mathrm{var}(X)=\lambda$$

We now list some important continuous distributions.

Beta

$$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0\le x\le 1; \qquad \alpha>0,\ \beta>0$$
$$\mathbb{E}(X)=\frac{\alpha}{\alpha+\beta}$$
$$\mathrm{var}(X)=\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

Cauchy

$$f(x)=\frac{1}{\pi(1+x^2)}, \qquad -\infty<x<\infty$$
$$\mathbb{E}(X) \text{ not defined}$$
$$\mathrm{var}(X)=\infty$$

Exponential

$$f(x)=\frac{1}{\theta}\exp\left(\frac{x}{\theta}\right), \qquad 0\le x<\infty; \qquad \theta>0$$
$$\mathbb{E}X=\theta$$
$$\mathrm{var}(X)=\theta^2$$

Logistic

$$f(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}, \qquad -\infty < x < \infty;$$
$$\mathbb{E}(X) = 0$$
$$\operatorname{var}(X) = \frac{\pi^2}{3}$$

Lognormal

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right), \qquad 0 \le x < \infty; \qquad \sigma > 0$$
$$\mathbb{E}(X) = \exp\left(\mu + \sigma^2/2\right)$$
$$\operatorname{var}(X) = \exp\left(2\mu + 2\sigma^2\right) - \exp\left(2\mu + \sigma^2\right)$$

Pareto

$$f(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \qquad \alpha \le x < \infty, \qquad \alpha > 0, \qquad \beta > 0$$
$$\mathbb{E}(X) = \frac{\beta\alpha}{\beta - 1}, \qquad \beta > 1$$
$$\operatorname{var}(X) = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \qquad \beta > 2$$

Uniform

$$f(x) = \frac{1}{b-a}, \qquad a \le x \le b$$
$$\mathbb{E}(X) = \frac{a+b}{2}$$
$$\operatorname{var}(X) = \frac{(b-a)^2}{12}$$

Weibull

$$f(x) = \frac{\gamma}{\beta} x^{\gamma-1} \exp\left(-\frac{x^\gamma}{\beta}\right), \qquad 0 \le x < \infty; \qquad \gamma > 0,\ \beta > 0$$

$$\mathbb{E}(X) = \beta^{1/\gamma} \Gamma\left(1 + \frac{1}{\gamma}\right)$$

$$\text{var}(X) = \beta^{2/\gamma} \left(\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right)\right)$$

Gamma

$$f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\theta}\right), \qquad 0 \le x < \infty; \qquad \alpha > 0,\ \theta > 0$$

$$\mathbb{E}(X) = \alpha\theta$$

$$\text{var}(X) = \alpha\theta^2$$

Chi-Square

$$f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} \exp\left(-\frac{x}{2}\right), \qquad 0 \le x < \infty; \qquad r > 0$$

$$\mathbb{E}(X) = r$$

$$\text{var}(X) = 2r$$

Normal

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty; \qquad -\infty < \mu < \infty,\ \sigma^2 > 0$$

$$\mathbb{E}(X) = \mu$$

$$\text{var}(X) = \sigma^2$$

Student t

$$f(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}, \qquad -\infty < x < \infty; \qquad r > 0$$

$$\mathbb{E}(X) = 0 \text{ if } r > 1$$

$$\text{var}(X) = \frac{r}{r-2} \text{ if } r > 2$$

B.6 Multivariate Random Variables

A pair of bivariate random variables (X, Y) is a function from the sample space into $\mathbb{R}^2$. The joint CDF of (X, Y) is

$$F(x, y) = \Pr(X \leq x, Y \leq y).$$

If F is continuous, the joint probability density function is

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

For a Borel measurable set $A \in R^2$,

$$\Pr((X, Y) \in A) = \int \int_A f(x, y) dx dy$$

For any measurable function $g(x, y)$,

$$\mathbb{E}g(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

The **marginal distribution** of X is

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) \\ &= \lim_{y \to \infty} F(x, y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, y) dy du \end{aligned}$$

so the **marginal density** of X is

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Similarly, the marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

The random variables X and Y are defined to be **independent** if $f(x,y) = f_X(x)f_Y(y)$. Furthermore, X and Y are independent if and only if there exist functions $g(x)$ and $h(y)$ such that $f(x,y) = g(x)h(y)$.

If X and Y are independent, then

$$\begin{aligned}\mathbb{E}\left(g(X)h(Y)\right) &= \int\int g(x)h(y)f(y,x)dydx \\ &= \int\int g(x)h(y)f_Y(y)f_X(x)dydx \\ &= \int g(x)f_X(x)dx \int h(y)f_Y(y)dy \\ &= \mathbb{E}\left(g\left(X\right)\right)\mathbb{E}\left(h\left(Y\right)\right). \end{aligned} \tag{B.10}$$

if the expectations exist. For example, if X and Y are independent then

$$\mathbb{E}(XY) = \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right).$$

Another implication of (B.10) is that if X and Y are independent and $Z = X + Y$, then

$$\begin{aligned} M_Z(\lambda) &= \mathbb{E}\left(\exp\left(\lambda\left(X+Y\right)\right)\right) \\ &= \mathbb{E}\left(\left(\exp\left(\lambda X\right)\exp\left(\lambda Y\right)\right)\right) \\ &= \mathbb{E}\left(\exp\left(\lambda' X\right)\right)\mathbb{E}\left(\exp\left(\lambda' Y\right)\right) \\ &= M_X(\lambda)M_Y(\lambda). \end{aligned} \tag{B.11}$$

The covariance between X and Y is

$$\operatorname{cov}(X,Y) = \sigma_{XY} = \mathbb{E}\left(\left(X - \mathbb{E}X\right)\left(Y - \mathbb{E}Y\right)\right) = \mathbb{E}\left(XY\right) - \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right).$$

The correlation between X and Y is

$$\operatorname{corr}\left(X,Y\right) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

The Cauchy-Schwarz Inequality implies that

$$|\rho_{XY}| \leq 1. \tag{B.12}$$

The correlation is a measure of linear dependence, free of units of measurement.

If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$. The reverse, however, is not true. For example, if $\mathbb{E}(X) = 0$ and $\mathbb{E}\left(X^3\right) = 0$, then $\operatorname{cov}(X, X^2) = 0$.

A useful fact is that

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y).$$

An implication is that if X and Y are independent, then

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y),$$

the variance of the sum is the sum of the variances.

A $k \times 1$ random vector $\boldsymbol{X} = (X_1, ..., X_k)'$ is a function from S to $\mathbb{R}^k$. Let $\boldsymbol{x} = (x_1, ..., x_k)'$ denote a vector in $\mathbb{R}^k$. (In this Appendix, we use bold to denote vectors. Bold capitals $\boldsymbol{X}$ are random vectors and bold lower case $\boldsymbol{x}$ are nonrandom vectors. Again, this is in distinction to the notation used in the bulk of the text) The vector $\boldsymbol{X}$ has the distribution and density functions

$$\begin{aligned} F(\boldsymbol{x}) &= \Pr(\boldsymbol{X} \leq \boldsymbol{x}) \\ f(\boldsymbol{x}) &= \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F(\boldsymbol{x}). \end{aligned}$$

For a measurable function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^s$, we define the expectation

$$\mathbb{E}\boldsymbol{g}(\boldsymbol{X}) = \int_{\mathbb{R}^k} g(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

where the symbol $d\boldsymbol{x}$ denotes $dx_1 \cdots dx_k$. In particular, we have the $k \times 1$ multivariate mean

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X})$$

and $k \times k$ covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbb{E}\left((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\right) \\ &= \mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}'\right) - \boldsymbol{\mu}\boldsymbol{\mu}' \end{aligned}$$

If the elements of $\boldsymbol{X}$ are mutually independent, then $\boldsymbol{\Sigma}$ is a diagonal matrix and

$$\operatorname{var}\left(\sum_{i=1}^{k} \boldsymbol{X}_i\right) = \sum_{i=1}^{k} \operatorname{var}(\boldsymbol{X}_i)$$

B.7 Conditional Distributions and Expectation

The **conditional density** of Y given $\boldsymbol{X} = \boldsymbol{x}$ is defined as

$$f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

if $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$. One way to derive this expression from the definition of conditional probability is

$$\begin{aligned}
f_{Y|X}(y \mid \boldsymbol{x}) &= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \Pr(Y \leq y \mid \boldsymbol{x} \leq \boldsymbol{X} \leq \boldsymbol{x} + \varepsilon) \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{\Pr(\{Y \leq y\} \cap \{\boldsymbol{x} \leq \boldsymbol{X} \leq \boldsymbol{x} + \varepsilon\})}{\Pr(\boldsymbol{x} \leq X \leq \boldsymbol{x} + \varepsilon)} \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{F(\boldsymbol{x} + \varepsilon, y) - F(\boldsymbol{x}, y)}{F_{\boldsymbol{X}}(\boldsymbol{x} + \varepsilon) - F_X(\boldsymbol{x})} \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{\frac{\partial}{\partial x} F(\boldsymbol{x} + \varepsilon, y)}{f_{\boldsymbol{X}}(\boldsymbol{x} + \varepsilon)} \\
&= \frac{\frac{\partial^2}{\partial x \partial y} F(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\
&= \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}.
\end{aligned}$$

The **conditional mean** or **conditional expectation** is the function

$$m(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x})\, dy.$$

The conditional mean $m(\boldsymbol{x})$ is a function, meaning that when $\boldsymbol{X}$ equals $\boldsymbol{x}$, then the expected value of Y is $m(\boldsymbol{x})$.

Similarly, we define the conditional variance of Y given $\boldsymbol{X} = \boldsymbol{x}$ as

$$\begin{aligned}
\sigma^2(\boldsymbol{x}) &= \operatorname{var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= \mathbb{E}\left((Y - m(\boldsymbol{x}))^2 \mid \boldsymbol{X} = \boldsymbol{x} \right) \\
&= \mathbb{E}\left(Y^2 \mid X = \boldsymbol{x} \right) - m(\boldsymbol{x})^2.
\end{aligned}$$

Evaluated at $\boldsymbol{x} = \boldsymbol{X}$, the conditional mean $m(\boldsymbol{X})$ and conditional variance $\sigma^2(\boldsymbol{X})$ are random variables, functions of $\boldsymbol{X}$. We write this as $\mathbb{E}(Y \mid \boldsymbol{X}) = m(\boldsymbol{X})$ and $\operatorname{var}(Y \mid \boldsymbol{X}) = \sigma^2(\boldsymbol{X})$. For example, if $\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \alpha + \beta'\boldsymbol{x}$, then $\mathbb{E}(Y \mid \boldsymbol{X}) = \alpha + \beta'\boldsymbol{X}$, a transformation of $\boldsymbol{X}$.

The following are important facts about conditional expectations.

Simple Law of Iterated Expectations:

$$\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X})) = \mathbb{E}(Y) \tag{B.13}$$

Proof:

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X})) &= \mathbb{E}(m(\boldsymbol{X})) \\
&= \int_{-\infty}^{\infty} m(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x}) dy d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(y, \boldsymbol{x})\, dy d\boldsymbol{x} \\
&= \mathbb{E}(Y).
\end{aligned}$$

Law of Iterated Expectations:

$$\mathbb{E}(\mathbb{E}(Y \mid \boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X}) = \mathbb{E}(Y \mid \boldsymbol{X}) \tag{B.14}$$

Conditioning Theorem. For any function $g(\boldsymbol{x})$,

$$\mathbb{E}(g(\boldsymbol{X})Y \mid \boldsymbol{X}) = g(\boldsymbol{X})\,\mathbb{E}(Y \mid \boldsymbol{X}) \tag{B.15}$$

Proof: Let

$$\begin{aligned}
h(\boldsymbol{x}) &= \mathbb{E}(g(\boldsymbol{X})Y \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= \int_{-\infty}^{\infty} g(\boldsymbol{x}) y f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x})\, dy \\
&= g(\boldsymbol{x}) \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x})\, dy \\
&= g(\boldsymbol{x}) m(\boldsymbol{x})
\end{aligned}$$

where $m(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$. Thus $h(\boldsymbol{X}) = g(\boldsymbol{X})m(\boldsymbol{X})$, which is the same as $\mathbb{E}(g(\boldsymbol{X})Y \mid \boldsymbol{X}) = g(\boldsymbol{X})\,\mathbb{E}(Y \mid \boldsymbol{X})$.

B.8 Transformations

Suppose that $\boldsymbol{X} \in \mathbb{R}^k$ with continuous distribution function $F_{\boldsymbol{X}}(\boldsymbol{x})$ and density $f_{\boldsymbol{X}}(\boldsymbol{x})$. Let $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ where $\boldsymbol{g}(\boldsymbol{x}) : \mathbb{R}^k \to \mathbb{R}^k$ is one-to-one, differentiable, and invertible. Let $\boldsymbol{h}(\boldsymbol{y})$ denote the inverse of $\boldsymbol{g}(\boldsymbol{x})$. The **Jacobian** is

$$J(\boldsymbol{y}) = \det\left(\frac{\partial}{\partial \boldsymbol{y}'}\boldsymbol{h}(\boldsymbol{y})\right).$$

Consider the univariate case $k = 1$. If $g(x)$ is an increasing function, then $g(X) \le Y$ if and only if $X \le h(Y)$, so the distribution function of Y is

$$\begin{aligned} F_Y(y) &= \Pr\left(g(X) \le y\right) \\ &= \Pr\left(X \le h(Y)\right) \\ &= F_X\left(h(Y)\right). \end{aligned}$$

Taking the derivative, the density of Y is

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

If $g(x)$ is a decreasing function, then $g(X) \le Y$ if and only if $X \ge h(Y)$, so

$$\begin{aligned} F_Y(y) &= \Pr\left(g(X) \le y\right) \\ &= 1 - \Pr\left(X \ge h(Y)\right) \\ &= 1 - F_X\left(h(Y)\right) \end{aligned}$$

and the density of Y is

$$f_Y(y) = -f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

We can write these two cases jointly as

$$f_Y(y) = f_X\left(h(Y)\right)\left|J(y)\right|. \tag{B.16}$$

This is known as the **change-of-variables** formula. This same formula (B.16) holds for $k > 1$, but its justification requires deeper results from analysis.

As one example, take the case $X \sim U[0,1]$ and $Y = -\log(X)$. Here, $g(x) = -\log(x)$ and $h(y) = \exp(-y)$ so the Jacobian is $J(y) = -\exp(-y)$. As the range of X is $[0,1]$, that for Y is $[0,\infty)$. Since $f_X(x) = 1$ for $0 \leq x \leq 1$ (B.16) shows that

$$f_Y(y) = \exp(-y), \qquad 0 \leq y \leq \infty,$$

an exponential density.

B.9 Inequalities

Jensen's Inequality. If $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is convex, then for any random vector $\boldsymbol{x}$ for which $\mathbb{E}\|\boldsymbol{x}\| < \infty$ and $\mathbb{E}|g(\boldsymbol{x})| < \infty$,

$$g(\mathbb{E}(\boldsymbol{x})) \leq \mathbb{E}(g(\boldsymbol{x})). \tag{B.17}$$

If $g(\cdot)$ concave, then the inequality is reversed.

Conditional Jensen's Inequality. If $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is convex, then for any random vectors $(\boldsymbol{y}, \boldsymbol{x})$ for which $\mathbb{E}\|\boldsymbol{y}\| < \infty$ and $\mathbb{E}\|g(\boldsymbol{y})\| < \infty$,

$$g(\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})) \leq \mathbb{E}(g(\boldsymbol{y}) \mid \boldsymbol{x}). \tag{B.18}$$

If $g(\cdot)$ concave, then the inequality is reversed.

Conditional Expectation Inequality. For any $r \geq 1$ such that $\mathbb{E}|y|^r < \infty$, then

$$\mathbb{E}(|\mathbb{E}(y \mid \boldsymbol{x})|^r) \leq \mathbb{E}(|y|^r) < \infty. \tag{B.19}$$

Expectation Inequality. For any random matrix $\boldsymbol{Y}$ for which $\mathbb{E}\|\boldsymbol{Y}\| < \infty$,

$$\|\mathbb{E}(\boldsymbol{Y})\| \leq \mathbb{E}\|\boldsymbol{Y}\|. \tag{B.20}$$

Hölder's Inequality. If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then for any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\mathbb{E}\left\|\boldsymbol{X}'\boldsymbol{Y}\right\| \leq \left(\mathbb{E}(\|\boldsymbol{X}\|^p)\right)^{1/p} \left(\mathbb{E}(\|\boldsymbol{Y}\|^q)\right)^{1/q}. \tag{B.21}$$

Cauchy-Schwarz Inequality. For any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\mathbb{E}\left\|\boldsymbol{X}'\boldsymbol{Y}\right\| \leq \left(\mathbb{E}\left(\left\|\boldsymbol{X}\right\|^2\right)\right)^{1/2}\left(\mathbb{E}\left(\left\|\boldsymbol{Y}\right\|^2\right)\right)^{1/2}. \tag{B.22}$$

Matrix Cauchy-Schwarz Inequality. Tripathi (1999). For any random $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^\ell$,

$$\mathbb{E}\left(\boldsymbol{y}\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-}\mathbb{E}\left(\boldsymbol{x}\boldsymbol{y}'\right) \leq \mathbb{E}\left(\boldsymbol{y}\boldsymbol{y}'\right) \tag{B.23}$$

Minkowski's Inequality. For any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\left(\mathbb{E}\left(\left\|\boldsymbol{X}+\boldsymbol{Y}\right\|^p\right)\right)^{1/p} \leq \left(\mathbb{E}\left(\left\|\boldsymbol{X}\right\|^p\right)\right)^{1/p} + \left(\mathbb{E}\left(\left\|\boldsymbol{Y}\right\|^p\right)\right)^{1/p} \tag{B.24}$$

Liapunov's Inequality. For any random $m \times n$ matrix $\boldsymbol{X}$ and $1 \leq r \leq p$,

$$\left(\mathbb{E}\left(\left\|\boldsymbol{X}\right\|^r\right)\right)^{1/r} \leq \left(\mathbb{E}\left(\left\|\boldsymbol{X}\right\|^p\right)\right)^{1/p} \tag{B.25}$$

Markov's Inequality (standard form). For any random vector $\boldsymbol{x}$ and non-negative function $g(\boldsymbol{x}) \geq 0$,

$$\Pr(g(\boldsymbol{x}) > \alpha) \leq \alpha^{-1}\mathbb{E}\left(g(\boldsymbol{x})\right). \tag{B.26}$$

Markov's Inequality (strong form). For any random vector $\boldsymbol{x}$ and non-negative function $g(\boldsymbol{x}) \geq 0$,

$$\Pr(g(\boldsymbol{x}) > \alpha) \leq \alpha^{-1}\mathbb{E}\left(g\left(\boldsymbol{x}\right)1\left(g(\boldsymbol{x}) > \alpha\right)\right). \tag{B.27}$$

Chebyshev's Inequality. For any random variable x,

$$\Pr(|x - \mathbb{E}x| > \alpha) \leq \frac{\operatorname{var}(x)}{\alpha^2}. \tag{B.28}$$

Proof of Jensen's Inequality (B.17). Since $g(\boldsymbol{u})$ is convex, at any point $\boldsymbol{u}$ there is a nonempty set of subderivatives (linear surfaces touching $g(\boldsymbol{u})$ at $\boldsymbol{u}$ but lying below $g(\boldsymbol{u})$ for all $\boldsymbol{u}$). Let $a + \boldsymbol{b}'\boldsymbol{u}$ be a subderivative of $g(\boldsymbol{u})$ at $\boldsymbol{u} = \mathbb{E}(\boldsymbol{x})$. Then for all $\boldsymbol{u}$, $g(\boldsymbol{u}) \geq a + \boldsymbol{b}'\boldsymbol{u}$ yet $g(\mathbb{E}(\boldsymbol{x})) = a + \boldsymbol{b}'\mathbb{E}(\boldsymbol{x})$. Applying expectations, $\mathbb{E}(g(\boldsymbol{x})) \geq a + \boldsymbol{b}'\mathbb{E}(\boldsymbol{x}) = g(\mathbb{E}(\boldsymbol{x}))$, as stated. ■

Proof of Conditional Jensen's Inequality. The same as the proof of (B.17), but using conditional expectations. The conditional expectations exist since $\mathbb{E}\|\boldsymbol{y}\| < \infty$ and $\mathbb{E}\|g(\boldsymbol{y})\| < \infty$. ■

Proof of Conditional Expectation Inequality. As the function $|u|^r$ is convex for $r \geq 1$, the Conditional Jensen's inequality implies

$$|\mathbb{E}(y \mid \boldsymbol{x})|^r \leq \mathbb{E}(|y|^r \mid \boldsymbol{x}).$$

Taking unconditional expectations and the law of iterated expectations, we obtain

$$\mathbb{E}(|\mathbb{E}(y \mid \boldsymbol{x})|^r) \leq \mathbb{E}(\mathbb{E}(|y|^r \mid \boldsymbol{x})) = \mathbb{E}(|y|^r) < \infty$$

as required. ■

Proof of Expectation Inequality. By the Triangle inequality, for $\lambda \in [0, 1]$,

$$\|\lambda \boldsymbol{U}_1 + (1 - \lambda)\boldsymbol{U}_2\| \leq \lambda \|\boldsymbol{U}_1\| + (1 - \lambda)\|\boldsymbol{U}_2\|$$

which shows that the matrix norm $g(\boldsymbol{U}) = \|\boldsymbol{U}\|$ is convex. Applying Jensen's Inequality (B.17) we find (B.20). ■

Proof of Hölder's Inequality. Since $\frac{1}{p} + \frac{1}{q} = 1$ an application of Jensen's Inequality (A.8) shows that for any real a and b

$$\exp\left[\frac{1}{p}a + \frac{1}{q}b\right] \leq \frac{1}{p}\exp(a) + \frac{1}{q}\exp(b).$$

Setting $u = \exp(a)$ and $v = \exp(b)$ this implies

$$u^{1/p}v^{1/q} \leq \frac{u}{p} + \frac{v}{q}$$

and this inequality holds for any $u > 0$ and $v > 0$.

Set $u = \|\boldsymbol{X}\|^p / \mathbb{E}(\|\boldsymbol{X}\|^p)$ and $v = \|\boldsymbol{Y}\|^q / \mathbb{E}(\|\boldsymbol{Y}\|^q)$. Note that $\mathbb{E}(u) = \mathbb{E}(v) = 1$. By the matrix Schwarz Inequality (A.20), $\|\boldsymbol{X}'\boldsymbol{Y}\| \leq \|\boldsymbol{X}\| \|\boldsymbol{Y}\|$. Thus

$$\begin{aligned}
\frac{\mathbb{E}\|\boldsymbol{X}'\boldsymbol{Y}\|}{(\mathbb{E}(\|\boldsymbol{X}\|^p))^{1/p}(\mathbb{E}(\|\boldsymbol{Y}\|^q))^{1/q}} &\leq \frac{\mathbb{E}(\|\boldsymbol{X}\|\|\boldsymbol{Y}\|)}{(\mathbb{E}(\|\boldsymbol{X}\|^p))^{1/p}(\mathbb{E}(\|\boldsymbol{Y}\|^q))^{1/q}} \\
&= \mathbb{E}\left(u^{1/p}v^{1/q}\right) \\
&\leq \mathbb{E}\left(\frac{u}{p} + \frac{v}{q}\right) \\
&= \frac{1}{p} + \frac{1}{q} \\
&= 1,
\end{aligned}$$

which is (B.21). ■

Proof of Cauchy-Schwarz Inequality. Special case of Hölder's with $p = q = 2$.

Proof of Matrix Cauchy-Schwarz Inequality. Define $e = \boldsymbol{y} - (\mathbb{E}(\boldsymbol{y}\boldsymbol{x}'))(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^-\boldsymbol{x}$. Note that $\mathbb{E}(\boldsymbol{e}\boldsymbol{e}') \geq 0$ is positive semi-definite. We can calculate that

$$\mathbb{E}(\boldsymbol{e}\boldsymbol{e}') = \mathbb{E}(\boldsymbol{y}\boldsymbol{y}') - (\mathbb{E}(\boldsymbol{y}\boldsymbol{x}'))(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^-\mathbb{E}(\boldsymbol{x}\boldsymbol{y}').$$

Since the left-hand-side is positive semi-definite, so is the right-hand-side, which means $\mathbb{E}(\boldsymbol{y}\boldsymbol{y}') \geq (\mathbb{E}(\boldsymbol{y}\boldsymbol{x}'))(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^-\mathbb{E}(\boldsymbol{x}\boldsymbol{y}')$ as stated. ■

Proof of Liapunov's Inequality. The function $g(u) = u^{p/r}$ is convex for $u > 0$ since $p \geq r$. Set $u = \|\boldsymbol{X}\|^r$. By Jensen's inequality, $g(\mathbb{E}(u)) \leq \mathbb{E}(g(u))$ or

$$(\mathbb{E}(\|\boldsymbol{X}\|^r))^{p/r} \leq \mathbb{E}\left((\|\boldsymbol{X}\|^r)^{p/r}\right) = \mathbb{E}(\|\boldsymbol{X}\|^p).$$

Raising both sides to the power $1/p$ yields $(\mathbb{E}(\|\boldsymbol{X}\|^r))^{1/r} \leq (\mathbb{E}(\|\boldsymbol{X}\|^p))^{1/p}$ as claimed. ■

Proof of Minkowski's Inequality. Note that by rewriting, using the triangle inequality (A.21), and then Hölder's Inequality to the two expectations

$$\begin{aligned}
\mathbb{E}\left(\|\boldsymbol{X}+\boldsymbol{Y}\|^{p}\right) &= \mathbb{E}\left(\|\boldsymbol{X}+\boldsymbol{Y}\|\,\|\boldsymbol{X}+\boldsymbol{Y}\|^{p-1}\right) \\
&\leq \mathbb{E}\left(\|\boldsymbol{X}\|\,\|\boldsymbol{X}+\boldsymbol{Y}\|^{p-1}\right)+\mathbb{E}\left(\|\boldsymbol{Y}\|\,\|\boldsymbol{X}+\boldsymbol{Y}\|^{p-1}\right) \\
&\leq \left(\mathbb{E}\left(\|\boldsymbol{X}\|^{p}\right)\right)^{1/p}\mathbb{E}\left(\left(\|\boldsymbol{X}+\boldsymbol{Y}\|^{q(p-1)}\right)^{1/q}\right) \\
&+ \left(\mathbb{E}\left(\|\boldsymbol{Y}\|^{p}\right)\right)^{1/p}\mathbb{E}\left(\left(\|\boldsymbol{X}+\boldsymbol{Y}\|^{q(p-1)}\right)^{1/q}\right) \\
&= \left(\left(\mathbb{E}\left(\|\boldsymbol{X}\|^{p}\right)\right)^{1/p}+\left(\mathbb{E}\left(\|\boldsymbol{Y}\|^{p}\right)\right)^{1/p}\right)\mathbb{E}\left(\left(\|\boldsymbol{X}+\boldsymbol{Y}\|^{p}\right)^{(p-1)/p}\right)
\end{aligned}$$

where the second equality picks q to satisfy $1/p+1/q=1$, and the final equality uses this fact to make the substitution $q=p/(p-1)$ and then collects terms. Dividing both sides by $\mathbb{E}\left(\left(\|\boldsymbol{X}+\boldsymbol{Y}\|^{p}\right)^{(p-1)/p}\right)$, we obtain (B.24). ■

Proof of Markov's Inequality. Let F denote the distribution function of $\boldsymbol{x}$. Then

$$\begin{aligned}
\Pr\left(g(\boldsymbol{x})\geq\alpha\right) &= \int_{\{g(\boldsymbol{u})\geq\alpha\}} dF(\boldsymbol{u}) \\
&\leq \int_{\{g(\boldsymbol{u})\geq\alpha\}} \frac{g(\boldsymbol{u})}{\alpha} dF(\boldsymbol{u}) \\
&= \alpha^{-1}\int 1\left(g(\boldsymbol{u})>\alpha\right) g(\boldsymbol{u}) dF(\boldsymbol{u}) \\
&= \alpha^{-1}\mathbb{E}\left(g\left(\boldsymbol{x}\right)1\left(g(\boldsymbol{x})>\alpha\right)\right)
\end{aligned}$$

the inequality using the region of integration $\{g(\boldsymbol{u})>\alpha\}$. This establishes the strong form (B.27). Since $1\left(g(\boldsymbol{x})>\alpha\right)\leq 1$, the final expression is less than $\alpha^{-1}\mathbb{E}\left(g(\boldsymbol{x})\right)$, establishing the standard form (B.26). ■

Proof of Chebyshev's Inequality. Define $y=(x-\mathbb{E}x)^2$ and note that $\mathbb{E}\left(y\right)=\operatorname{var}\left(x\right)$. The events $\{|x-\mathbb{E}x|>\alpha\}$ and $\left\{y>\alpha^2\right\}$ are equal, so by an application Markov's inequality we find

$$\Pr(|x-\mathbb{E}x|>\alpha)=\Pr(y>\alpha^2)\leq\alpha^{-2}\mathbb{E}\left(y\right)=\alpha^{-2}\operatorname{var}\left(x\right)$$

as stated. ■

Appendix C

Numerical Optimization

Many econometric estimators are defined by an optimization problem of the form

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, Q(\boldsymbol{\theta}) \tag{C.1}$$

where the parameter is $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ and the criterion function is $Q(\boldsymbol{\theta}) : \boldsymbol{\Theta} \to \mathbb{R}$. For example NLLS, GLS, MLE and GMM estimators take this form. In most cases, $Q(\boldsymbol{\theta})$ can be computed for given $\boldsymbol{\theta}$, but $\widehat{\boldsymbol{\theta}}$ is not available in closed form. In this case, numerical methods are required to obtain $\hat{\boldsymbol{\theta}}$.

C.1 Grid Search

Many optimization problems are either one dimensional ($m = 1$) or involve one-dimensional optimization as a sub-problem (for example, a line search). In this context grid search may be employed.

Grid Search. Let $\Theta = [a, b]$ be an interval. Pick some $\varepsilon > 0$ and set $G = (b - a)/\varepsilon$ to be the number of gridpoints. Construct an equally spaced grid on the region $[a, b]$ with G gridpoints, which is $\{\boldsymbol{\theta}(j) = a + j(b - a)/G : j = 0, ..., G\}$. At each point evaluate the criterion function and find the gridpoint which yields the smallest value of the criterion, which is $\boldsymbol{\theta}(\hat{\jmath})$ where $\hat{\jmath} = \operatorname{argmin}_{0\le j\le G} Q(\boldsymbol{\theta}(j))$. This value $\boldsymbol{\theta}(\hat{\jmath})$ is the gridpoint estimate of $\widehat{\boldsymbol{\theta}}$**.** If the grid is sufficiently fine to capture small oscillations in $Q(\boldsymbol{\theta})$, the approximation error is bounded by ε, that is, $\left|\boldsymbol{\theta}(\hat{\jmath}) - \widehat{\boldsymbol{\theta}}\right| \le \varepsilon$. Plots of $Q(\boldsymbol{\theta}(j))$ against $\boldsymbol{\theta}(j)$

can help diagnose errors in grid selection. This method is quite robust but potentially costly.

Two-Step Grid Search. The gridsearch method can be refined by a two-step execution. For an error bound of ε pick G so that $G^2 = (b-a)/\varepsilon$ For the first step define an equally spaced grid on the region $[a, b]$ with G gridpoints, which is $\{\boldsymbol{\theta}(j) = a + j(b-a)/G : j = 0, ..., G\}$. At each point evaluate the criterion function and let $\hat{\jmath} = \text{argmin}_{0 \le j \le G} Q(\boldsymbol{\theta}(j))$. For the second step define an equally spaced grid on $[\boldsymbol{\theta}(\hat{\jmath}-1), \boldsymbol{\theta}(\hat{\jmath}+1)]$ with G gridpoints, which is $\{\boldsymbol{\theta}'(k) = \boldsymbol{\theta}(\hat{\jmath}-1) + 2k(b-a)/G^2 : k = 0, ..., G\}$. Let $\hat{k} = \text{argmin}_{0 \le k \le G} Q(\boldsymbol{\theta}'(k))$. The estimate of $\widehat{\boldsymbol{\theta}}$ is $\boldsymbol{\theta}\left(\hat{k}\right)$. The advantage of the two-step method over a one-step grid search is that the number of function evaluations has been reduced from $(b-a)/\varepsilon$ to $2\sqrt{(b-a)/\varepsilon}$ which can be substantial. The disadvantage is that if the function $Q(\boldsymbol{\theta})$ is irregular, the first-step grid may not bracket $\hat{\boldsymbol{\theta}}$ which thus would be missed.

C.2 Gradient Methods

Gradient Methods are iterative methods which produce a sequence $\boldsymbol{\theta}_i : i = 1, 2, ...$ which are designed to converge to $\hat{\boldsymbol{\theta}}$. All require the choice of a starting value $\boldsymbol{\theta}_1$, and all require the computation of the **gradient** of $Q(\boldsymbol{\theta})$

$$\boldsymbol{g}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta})$$

and some require the **Hessian**

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} Q(\boldsymbol{\theta}).$$

If the functions $\boldsymbol{g}(\boldsymbol{\theta})$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$ are not analytically available, they can be calculated numerically. Take the $j'th$ element of $\boldsymbol{g}(\boldsymbol{\theta})$. Let δ_j be the $j'th$ unit vector (zeros everywhere except for a one in the $j'th$ row). Then for ε small

$$g_j(\boldsymbol{\theta}) \simeq \frac{Q(\boldsymbol{\theta} + \delta_j \varepsilon) - Q(\boldsymbol{\theta})}{\varepsilon}.$$

Similarly,

$$g_{jk}(\boldsymbol{\theta}) \simeq \frac{Q(\boldsymbol{\theta} + \delta_j \varepsilon + \delta_k \varepsilon) - Q(\boldsymbol{\theta} + \delta_k \varepsilon) - Q(\boldsymbol{\theta} + \delta_j \varepsilon) + Q(\boldsymbol{\theta})}{\varepsilon^2}$$

In many cases, numerical derivatives can work well but can be computationally costly relative to analytic derivatives. In some cases, however, numerical derivatives can be quite unstable.

Most gradient methods are a variant of **Newton's method** which is based on a quadratic approximation. By a Taylor's expansion for $\boldsymbol{\theta}$ close to $\hat{\boldsymbol{\theta}}$

$$0 = \boldsymbol{g}(\widehat{\boldsymbol{\theta}}) \simeq \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$$

which implies

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta})^{-1} \boldsymbol{g}(\boldsymbol{\theta}).$$

This suggests the iteration rule

$$\widehat{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)^{-1} \boldsymbol{g}(\boldsymbol{\theta}_i).$$

where

One problem with Newton's method is that it will send the iterations in the wrong direction if $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)$ is not positive definite. One modification to prevent this possibility is quadratic hill-climbing which sets

$$\widehat{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_i - \left(\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i) + \alpha_i \boldsymbol{I}_m\right)^{-1} \boldsymbol{g}(\boldsymbol{\theta}_i).$$

where α_i is set just above the smallest eigenvalue of $\boldsymbol{H}(\boldsymbol{\theta}_i)$ if $\boldsymbol{H}(\boldsymbol{\theta})$ is not positive definite.

Another productive modification is to add a scalar **steplength** λ_i. In this case the iteration rule takes the form

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{D}_i \boldsymbol{g}_i \lambda_i \tag{C.2}$$

where $\boldsymbol{g}_i = \boldsymbol{g}(\boldsymbol{\theta}_i)$ and $\boldsymbol{D}_i = \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)^{-1}$ for Newton's method and $D_i = (\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i) + \alpha_i \boldsymbol{I}_m)$ for quadratic hill-climbing.

Allowing the steplength to be a free parameter allows for a line search, a one-dimensional optimization. To pick λ_i write the criterion function as a function of λ

$$Q(\lambda) = Q(\boldsymbol{\theta}_i + \boldsymbol{D}_i \boldsymbol{g}_i \lambda)$$

a one-dimensional optimization problem. There are two common methods to perform a line search. A **quadratic approximation** evaluates the first

and second derivatives of $Q(\lambda)$ with respect to λ, and picks λ_i as the value minimizing this approximation. The **half-step** method considers the sequence $\lambda = 1,\ 1/2,\ 1/4,\ 1/8,\ \ldots$. Each value in the sequence is considered and the criterion $Q(\boldsymbol{\theta}_i + \boldsymbol{D}_i \boldsymbol{g}_i \lambda)$ evaluated. If the criterion has improved over $Q(\boldsymbol{\theta}_i)$, use this value, otherwise move to the next element in the sequence.

Newton's method does not perform well if $Q(\boldsymbol{\theta})$ is irregular, and it can be quite computationally costly if $\boldsymbol{H}(\boldsymbol{\theta})$ is not analytically available. These problems have motivated alternative choices for the weight matrix D_i. These methods are called **Quasi-Newton** methods. Two popular methods are do to Davidson-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS).

Let

$$\begin{aligned}\Delta \boldsymbol{g}_i &= \boldsymbol{g}_i - \boldsymbol{g}_{i-1} \\ \Delta \boldsymbol{\theta}_i &= \boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\end{aligned}$$

and . The DFP method sets

$$\boldsymbol{D}_i = \boldsymbol{D}_{i-1} + \frac{\Delta \boldsymbol{\theta}_i \Delta \boldsymbol{\theta}_i'}{\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i} + \frac{\boldsymbol{D}_{i-1} \Delta \boldsymbol{g}_i \Delta \boldsymbol{g}_i' \boldsymbol{D}_{i-1}}{\Delta \boldsymbol{g}_i' \boldsymbol{D}_{i-1} \Delta \boldsymbol{g}_i}.$$

The BFGS methods sets

$$\boldsymbol{D}_i = \boldsymbol{D}_{i-1} + \frac{\Delta \boldsymbol{\theta}_i \Delta \boldsymbol{\theta}_i'}{\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i} - \frac{\Delta \boldsymbol{\theta}_i \Delta \boldsymbol{\theta}_i'}{\left(\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i\right)^2} \Delta g_i' \boldsymbol{D}_{i-1} \Delta \boldsymbol{g}_i + \frac{\Delta \boldsymbol{\theta}_i \Delta \boldsymbol{g}_i' \boldsymbol{D}_{i-1}}{\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i} + \frac{\boldsymbol{D}_{i-1} \Delta \boldsymbol{g}_i \Delta \boldsymbol{\theta}_i'}{\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i}.$$

For any of the gradient methods, the iterations continue until the sequence has converged in some sense. This can be defined by examining whether $|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}|$, $|Q(\boldsymbol{\theta}_i) - Q(\boldsymbol{\theta}_{i-1})|$ or $|g(\boldsymbol{\theta}_i)|$ has become small.

C.3 Derivative-Free Methods

All gradient methods can be quite poor in locating the global minimum when $Q(\boldsymbol{\theta})$ has several local minima. Furthermore, the methods are not well defined when $Q(\boldsymbol{\theta})$ is non-differentiable. In these cases, alternative optimization methods are required. One example is the **simplex method** of Nelder-Mead (1965).

A more recent innovation is the method of **simulated annealing (SA).** For a review see Goffe, Ferrier, and Rodgers (1994). The SA method is a sophisticated random search. Like the gradient methods, it relies on an iterative sequence. At each iteration, a random variable is drawn and added to the current value of the parameter. If the resulting criterion is decreased, this new value is accepted. If the criterion is increased, it may still be accepted depending on the extent of the increase and another randomization. The latter property is needed to keep the algorithm from selecting a local minimum. As the iterations continue, the variance of the random innovations is shrunk. The SA algorithm stops when a large number of iterations is unable to improve the criterion. The SA method has been found to be successful at locating global minima. The downside is that it can take considerable computer time to execute.

Bibliography

[1] Abadir, Karim M. and Jan R. Magnus (2005): *Matrix Algebra*, Cambridge University Press.

[2] Aitken, A.C. (1935): "On least squares and linear combinations of observations," *Proceedings of the Royal Statistical Society*, 55, 42-48.

[3] Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle." In B. Petroc and F. Csake, eds., *Second International Symposium on Information Theory.*

[4] Anderson, T.W. and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics*, 20, 46-63.

[5] Andrews, Donald W. K. (1988): "Laws of large numbers for dependent non-identically distributed random variables,' *Econometric Theory,* 4, 458-467.

[6] Andrews, Donald W. K. (1991), "Asymptotic normality of series estimators for nonparameric and semiparametric regression models," *Econometrica*, 59, 307-345.

[7] Andrews, Donald W. K. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821-8516.

[8] Andrews, Donald W. K. and Moshe Buchinsky: (2000): "A three-step method for choosing the number of bootstrap replications," *Econometrica*, 68, 23-51.

[9] Andrews, Donald W. K. and Werner Ploberger (1994): "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica,* 62, 1383-1414.

[10] Ash, Robert B. (1972): *Real Analysis and Probability*, Academic Press.

[11] Basmann, R. L. (1957): "A generalized classical method of linear estimation of coefficients in a structural equation," *Econometrica,* 25, 77-83.

[12] Bekker, P.A. (1994): "Alternative approximations to the distributions of instrumental variable estimators, *Econometrica,* 62, 657-681.

[13] Billingsley, Patrick (1968): *Convergence of Probability Measures.* New York: Wiley.

[14] Billingsley, Patrick (1995): *Probability and Measure,* 3rd Edition, New York: Wiley.

[15] Bose, A. (1988): "Edgeworth correction by bootstrap in autoregressions," *Annals of Statistics*, 16, 1709-1722.

[16] Box, George E. P. and Dennis R. Cox, (1964). "An analysis of transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211-252.

[17] Breusch, T.S. and A.R. Pagan (1979): "The Lagrange multiplier test and its application to model specification in econometrics," *Review of Economic Studies*, 47, 239-253.

[18] Brown, B. W. and Whitney K. Newey (2002): "GMM, efficient bootstrapping, and improved inference ," *Journal of Business and Economic Statistics.*

[19] Card, David (1995): "Using geographic variation in college proximity to estimate the return to schooling," in *Aspects of Labor Market Behavior: Essays in Honour of John Vanderkamp*, L.N. Christofides, E.K. Grant, and R. Swidinsky, editors. Toronto: University of Toronto Press.

[20] Carlstein, E. (1986): "The use of subseries methods for estimating the variance of a general statistic from a stationary time series," *Annals of Statistics*, 14, 1171-1179.

[21] Casella, George and Roger L. Berger (2002): *Statistical Inference*, 2nd Edition, Duxbury Press.

[22] Chamberlain, Gary (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.

[23] Choi, In and Peter C.B. Phillips (1992): "Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations," *Journal of Econometrics*, 51, 113-150.

[24] Chow, G.C. (1960): "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, 28, 591-603.

[25] Cragg, John (1992): "Quasi-Aitken Estimation for Heterskedasticity of Unknown Form" *Journal of Econometrics*, 54, 179-201.

[26] Davidson, James (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.

[27] Davison, A.C. and D.V. Hinkley (1997): *Bootstrap Methods and their Application*. Cambridge University Press.

[28] Dickey, D.A. and W.A. Fuller (1979): "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, 74, 427-431.

[29] Donald Stephen G. and Whitney K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.

[30] Dufour, J.M. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365-1387.

[31] Efron, Bradley (1979): "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1-26.

[32] Efron, Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

[33] Efron, Bradley and R.J. Tibshirani (1993): *An Introduction to the Bootstrap*, New York: Chapman-Hall.

[34] Eicker, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34, 447-456.

[35] Engle, Robert F. and Clive W. J. Granger (1987): "Co-integration and error correction: Representation, estimation and testing," *Econometrica*, 55, 251-276.

[36] Frisch, Ragnar (1933): "Editorial," *Econometrica*, 1, 1-4.

[37] Frisch, Ragnar and F. Waugh (1933): "Partial time regressions as compared with individual trends," *Econometrica*, 1, 387-401.

[38] Gallant, A. Ronald and D.W. Nychka (1987): "Seminonparametric maximum likelihood estimation," *Econometrica*, 55, 363-390.

[39] Gallant, A. Ronald and Halbert White (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. New York: Basil Blackwell.

[40] Galton, Francis (1886): "Regression Towards Mediocrity in Hereditary Stature," The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

[41] Goldberger, Arthur S. (1964): *Econometric Theory*, Wiley.

[42] Goldberger, Arthur S. (1968): *Topics in Regression Analysis*, Macmillan

[43] Goldberger, Arthur S. (1991): *A Course in Econometrics*. Cambridge: Harvard University Press.

[44] Goffe, W.L., G.D. Ferrier and J. Rogers (1994): "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, 60, 65-99.

[45] Gosset, William S. (a.k.a. "Student") (1908): "The probable error of a mean," *Biometrika*, 6, 1-25.

[46] Gauss, K.F. (1809): "Theoria motus corporum coelestium," in *Werke*, Vol. VII, 240-254.

[47] Granger, Clive W. J. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424-438.

[48] Granger, Clive W. J. (1981): "Some properties of time series data and their use in econometric specification," *Journal of Econometrics*, 16, 121-130.

[49] Granger, Clive W. J. and Timo Teräsvirta (1993): *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

[50] Gregory, A. and M. Veall (1985): "On formulating Wald tests of nonlinear restrictions," *Econometrica*, 53, 1465-1468,

[51] Haavelmo, T. (1944): "The probability approach in econometrics," *Econometrica*, supplement, 12.

[52] Hall, A. R. (2000): "Covariance matrix estimation and the power of the overidentifying restrictions test," *Econometrica*, 68, 1517-1527,

[53] Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

[54] Hall, P. (1994): "Methodology and theory for the bootstrap," *Handbook of Econometrics, Vol. IV*, eds. R.F. Engle and D.L. McFadden. New York: Elsevier Science.

[55] Hall, P. and J.L. Horowitz (1996): "Bootstrap critical values for tests based on Generalized-Method-of-Moments estimation," *Econometrica*, 64, 891-916.

[56] Hahn, J. (1996): "A note on bootstrapping generalized method of moments estimators," *Econometric Theory*, 12, 187-197.

[57] Hamilton, James D. (1994) *Time Series Analysis*.

[58] Hansen, Bruce E. (1992): "Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends," *Journal of Econometrics*, 53, 87-121.

[59] Hansen, Bruce E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413-430.

[60] Hansen, Bruce E. (2006): "Edgeworth expansions for the Wald and GMM statistics for nonlinear restrictions," *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, edited by Dean Corbae, Steven N. Durlauf and Bruce E. Hansen. Cambridge University Press.

[61] Hansen, Lars Peter (1982): "Large sample properties of generalized method of moments estimators, *Econometrica*, 50, 1029-1054.

[62] Hansen, Lars Peter, John Heaton, and A. Yaron (1996): "Finite sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, 14, 262-280.

[63] Hausman, J.A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1271.

[64] Heckman, J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153-161.

[65] Horn, S.D., R.A. Horn, and D.B. Duncan. (1975) "Estimating heteroscedastic variances in linear model," *Journal of the American Statistical Association*, 70, 380-385.

[66] Horowitz, Joel (2001): "The Bootstrap," *Handbook of Econometrics, Vol. 5*, J.J. Heckman and E.E. Leamer, eds., Elsevier Science, 3159-3228.

[67] Imbens, G.W. (1997): "One step estimators for over-identified generalized method of moments models," Review of Economic Studies, 64, 359-383.

[68] Imbens, G.W., R.H. Spady and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333-357.

[69] Jarque, C.M. and A.K. Bera (1980): "Efficient tests for normality, homoskedasticity and serial independence of regression residuals, *Economic Letters*, 6, 255-259.

[70] Johansen, S. (1988): "Statistical analysis of cointegrating vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.

[71] Johansen, S. (1991): "Estimation and hypothesis testing of cointegration vectors in the presence of linear trend," *Econometrica*, 59, 1551-1580.

[72] Johansen, S. (1995): *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*, Oxford University Press.

[73] Johansen, S. and K. Juselius (1992): "Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for the UK," *Journal of Econometrics*, 53, 211-244.

[74] Kitamura, Y. (2001): "Asymptotic optimality and empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661-1672.

[75] Kitamura, Y. and M. Stutzer (1997): "An information-theoretic alternative to generalized method of moments," *Econometrica*, 65, 861-874..

[76] Koenker, Roger (2005): *Quantile Regression*. Cambridge University Press.

[77] Kunsch, H.R. (1989): "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217-1241.

[78] Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of Econometrics*, 54, 159-178.

[79] Lafontaine, F. and K.J. White (1986): "Obtaining any Wald statistic you want," *Economics Letters*, 21, 35-40.

[80] Lehmann, E.L. and George Casella (1998): *Theory of Point Estimation*, 2nd Edition, Springer.

[81] Lehmann, E.L. and Joseph P. Romano (2005): *Testing Statistical Hypotheses*, 3rd Edition, Springer.

[82] Lindeberg, Jarl Waldemar, (1922): "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, 15, 211-225.

[83] Li, Qi and Jeffrey Racine (2007) *Nonparametric Econometrics*.

[84] Lovell, M.C. (1963): "Seasonal adjustment of economic time series," *Journal of the American Statistical Association*, 58, 993-1010.

[85] MacKinnon, James G. (1990): "Critical values for cointegration," in Engle, R.F. and C.W. Granger (eds.) *Long-Run Economic Relationships: Readings in Cointegration*, Oxford, Oxford University Press.

[86] MacKinnon, James G. and Halbert White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305-325.

[87] Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley and Sons.

[88] Mann, H.B. and A. Wald (1943). "On stochastic limit and order relationships," *The Annals of Mathematical Statistics* 14, 217–226.

[89] Muirhead, R.J. (1982): *Aspects of Multivariate Statistical Theory*. New York: Wiley.

[90] Nelder, J. and R. Mead (1965): "A simplex method for function minimization," *Computer Journal*, 7, 308-313.

[91] Nerlove, Marc (1963): "Returns to Scale in Electricity Supply," Chapter 7 of *Measurement in Economics* (C. Christ, et al, eds.). Stanford: Stanford University Press, 167-198.

[92] Newey, Whitney K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99-135.

[93] Newey, Whitney K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147-168.

[94] Newey, Whitney K. and Daniel L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Robert Engle and Daniel McFadden, (eds.) *Handbook of Econometrics*, vol. IV, 2111-2245, North Holland: Amsterdam.

[95] Newey, Whitney K. and Kenneth D. West (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.

[96] Owen, Art B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237-249.

[97] Owen, Art B. (2001): *Empirical Likelihood*. New York: Chapman & Hall.

[98] Park, Joon Y. and Peter C. B. Phillips (1988): "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, 56, 1065-1083,

[99] Phillips, Peter C.B. (1989): "Partially identified econometric models," *Econometric Theory*, 5, 181-240.

[100] Phillips, Peter C.B. and Sam Ouliaris (1990): "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 58, 165-193.

[101] Politis, D.N. and J.P. Romano (1996): "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.

[102] Potscher, B.M. (1991): "Effects of model selection on inference," *Econometric Theory*, 7, 163-185.

[103] Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300-325.

[104] Ramsey, J. B. (1969): "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society*, Series B, 31, 350-371.

[105] Rudin, W. (1987): *Real and Complex Analysis*, 3rd edition. New York: McGraw-Hill.

[106] Runge, Carl (1901): "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten," *Zeitschrift für Mathematik und Physik*, 46, 224-243.

[107] Said, S.E. and D.A. Dickey (1984): "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, 71, 599-608.

[108] Secrist, Horace (1933): *The Triumph of Mediocrity in Business*. Evanston: Northwestern University.

[109] Shao, J. and D. Tu (1995): *The Jackknife and Bootstrap*. NY: Springer.

[110] Sargan, J.D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica, 2*6, 393-415.

[111] Shao, Jun (2003): *Mathematical Statistics*, 2nd edition, Springer.

[112] Sheather, S.J. and M.C. Jones (1991): "A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

[113] Shin, Y. (1994): "A residual-based test of the null of cointegration against the alternative of no cointegration," *Econometric Theory*, 10, 91-115.

[114] Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[115] Sims, C.A. (1972): "Money, income and causality," *American Economic Review*, 62, 540-552.

[116] Sims, C.A. (1980): "Macroeconomics and reality," *Econometrica, 48*, 1-48.

[117] Staiger, D. and James H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica,* 65, 557-586.

[118] Stock, James H. (1987): "Asymptotic properties of least squares estimators of cointegrating vectors," *Econometrica,* 55, 1035-1056.

[119] Stock, James H. (1991): "Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series," *Journal of Monetary Economics,* 28, 435-460.

[120] Stock, James H. and Jonathan H. Wright (2000): "GMM with weak identification," *Econometrica,* 68, 1055-1096.

[121] Stock, James H. and Mark W. Watson (2010): *Introduction to Econometrics*, 3rd edition, Addison-Wesley.

[122] Stone, Marshall H. (1937): "Applications of the Theory of Boolean Rings to General Topology," *Transactions of the American Mathematical Society*, 41, 375-481.

[123] Stone, Marshall H. (1948): "The Generalized Weierstrass Approximation Theorem," *Mathematics Magazine*, 21, 167-184.

[124] Theil, Henri. (1953): "Repeated least squares applied to complete equation systems," The Hague, Central Planning Bureau, mimeo.

[125] Theil, Henri (1961): *Economic Forecasts and Policy*. Amsterdam: North Holland.

[126] Theil, Henri. (1971): *Principles of Econometrics*, New York: Wiley.

[127] Tobin, James (1958): "Estimation of relationships for limited dependent variables," *Econometrica, 2*6, 24-36.

[128] Tripathi, Gautam (1999): "A matrix extension of the Cauchy-Schwarz inequality," Economics Letters, 63, 1-3.

[129] van der Vaart, A.W. (1998): *Asymptotic Statistics*, Cambridge University Press.

[130] Wald, A. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, 54, 426-482.

[131] Wang, J. and E. Zivot (1998): "Inference on structural parameters in instrumental variables regression with weak instruments," *Econometrica,* 66, 1389-1404.

[132] Weierstrass, K. (1885): "Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen," *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, 1885.

[133] White, Halbert (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.

[134] White, Halbert (1984): *Asymptotic Theory for Econometricians,* Academic Press.

[135] Wooldridge, Jeffrey M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, MIT Press.

[136] Wooldridge, Jeffrey M. (2009) *Introductory Econometrics: A Modern Approach,* 4th edition, Southwestern

[137] Zellner, Arnold. (1962): "An efficient method of estimating seemingly unrelated regressions, and tests for aggregation bias," *Journal of the American Statistical Association*, 57, 348-368.

[138] Zhang, Fuzhen and Qingling Zhang (2006): "Eigenvalue inequalities for matrix product," *IEEE Transactions on Automatic Control*, 51, 1506-1509.)