Part IV

The Appendices

Appendix A

Mathematical and Statistical Topics

A.1 Trigonometry

Trigonometry is a branch of mathematics that studies triangles. The study of trigonometry deals with the relationships between the lengths of the sides and measures of the angles of a right triangle. As we will see, in two-dimensions, trigonometry can also be used to describe the relationships dealing with angles and the motion and shape of waves. This tutorial begins with the most basic of definitions. Although basic, the quantities defined below are fundamental to trigonometry.

Definition A.1.1 Right Triangle.
If the measures of one of the angles of a triangle is $\pi/2$ ($90°$), then the triangle is a **right triangle**. *The side of the triangle that is opposite the right angle is the* **hypotenuse** *of the right triangle. The length of the hypotenuse will be denoted by r.*

The *triangle postulate* is the statement that the sum of the measures of all the angles of a triangle is equal to π ($180°$). So for a right triangle, the two non-right angles must have measures that are less than $\pi/2$.

For any right triangle, the relationships between the lengths of the sides and the measures of the angles have specific definitions. These functions are the basic building blocks of trigonometry.

Definition A.1.2 Basic Trigonometric Functions.
Let θ denote the measure of an angle of a right triangle. Let θ denote an angle of a right triangle, a denote the length of the side opposite θ, and b denote the length of the side adjacent to θ. Then for a right triangle, the trigonometric functions **sine**, **cosine**, *and* **tangent** *are defined in terms of the ratios of the lengths of sides adjacent and opposite θ and the length of the hypotenuse.*

$$\sin\theta = \frac{\textit{length of opposite side}}{\textit{length of hypotenuse}} = \frac{a}{r} \tag{A.1}$$

$$\cos\theta = \frac{\textit{length of adjacent side}}{\textit{length of hypotenuse}} = \frac{b}{r} \tag{A.2}$$

$$\tan\theta \;=\; \frac{\textit{length of opposite side}}{\textit{length of adjacent side}} = \frac{a}{b} = \frac{\sin\theta}{\cos\theta}. \tag{A.3}$$

For an illustration of these definitions, see Figure A.1.

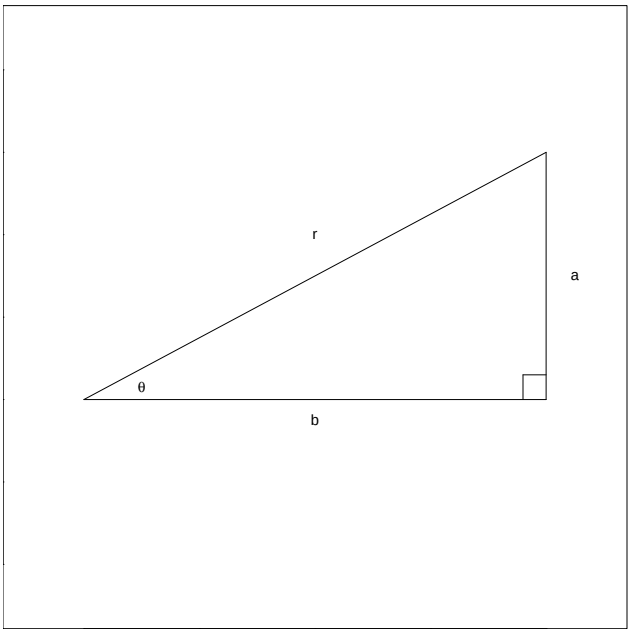


Figure A.1: Illustration of relationships of sides of triangle to angle A as described in Definition A.1.1.

The famous Pythagorean Theorem says that for a right triangle,

$$a^2 + b^2 = r^2.$$

Definition A.1.3 Reciprocals of Basic Trignometric Functions.
The reciprocals of the sine, cosine, and tangent functions are called **secant**, **cosecant**, *and* **cotangent**, *respectively; that is,*

$$\sec\theta = \frac{1}{\cos\theta}, \quad \csc\theta = \frac{1}{\sin\theta}, \quad \textit{and} \quad \cot\theta = \frac{1}{\tan\theta}. \tag{A.4}$$

Figure A.2 displays sine and cosine as a function of θ for $0 \le \theta \le 6\pi$. In this figure, the sine and cosine function go through three cycles. Examination of this graph illustrate that the sine and cosine function are periodic, having period 2π, as is the tangent function, which is not included in the graph. Mathematically, this is written as

$$\sin\theta = \sin(\theta + 2\pi), \quad \cos\theta = \cos(\theta + 2\pi), \quad \text{and} \quad \tan\theta = \tan(\theta + 2\pi).$$

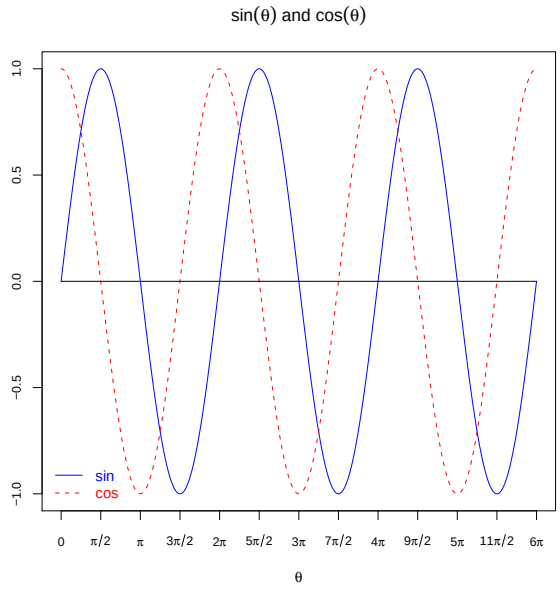


Figure A.2: Graphs of cosine and sine as functions of θ.

Figure A.2 also readily illustrate how, in two-dimensions, trigonometry can also be used to describe the relationships dealing with angles and the motion and shape of waves.

For $0 \leq \theta \leq \pi/2$, it is also the case that

$$\begin{aligned} \sin(\theta) &= -\sin(\theta+\pi) = -\sin(\theta-\pi) = \sin(2\pi-\theta), \\ \cos(\theta) &= -\cos(\theta+\pi) = -\cos(\theta-\pi) = \cos(2\pi-\theta) \end{aligned} \tag{A.5}$$

but

$$\tan(\theta) = \tan(\theta+\pi) = \tan(\theta-\pi).$$

Figure A.3 illustrates this relationship for the sine function. Because the sine, cosine, and tangent functions are periodic, then, strictly speaking, they do not have a mathematical inverse. To define their inverse, it is necessary to restrict the values of θ so that the functions are one-to-one and onto; that is, are bijective. Additionally, for the secant, cosecant, and cotangent, the value of θ must not be so that the denominator is zero.

Definition A.1.4 Inverse of Basic Trigonometric Functions.
For restricted domain of the trigonometric functions, we may define their inverses in the following manner.

(a) For $-\pi/2 \leq \theta \leq \pi/2$, the inverse of the sine function is called the **arcsine** *and is defined by*

$$\theta = \arcsin\phi \quad \textit{if} \quad \phi = \sin\theta.$$

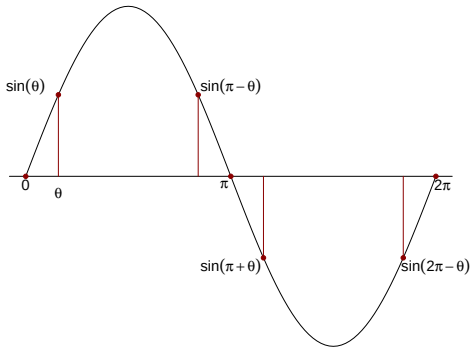


Figure A.3: Sine function for $0 \leq \theta \leq 2\pi$ illustrating equation (A.5).

(b) For $0 \leq \theta \leq \pi$, the inverse of the cosine function is called the **arccosine** *and is defined by*

$$\theta = \arccos\phi \quad \textit{if} \quad \phi = \cos\theta.$$

(c) For $-\pi/2 \leq \theta \leq \pi/2$, the inverse of the tangent function is called the **arctangent** *and is defined by*

$$\theta = \arctan\phi \quad \textit{if} \quad \phi = \tan\theta. \tag{A.6}$$

(d) For $-\pi/2 \leq \theta \leq \pi/2$ and $\theta \neq 0$, the inverse of the cosecant is called the **arccosecant** *and is defined by*

$$\theta = arccsc\phi \quad \textit{if} \quad \phi = \csc\theta.$$

(e) For $0 \leq \theta \leq \pi$ and $\theta \neq \pi/2$, the inverse of the secant is called the **arcsecant** *and is defined by*

$$\theta = arcsec\phi \quad \textit{if} \quad \phi = \sec\theta.$$

(f) For $0 \leq \theta \leq \pi$, the inverse of the cotangent is called the **arccotangent** *and is defined by*

$$\theta = arccot\phi \quad \textit{if} \quad \phi = \cot\theta.$$

Equation (1.1) in Theoretical Exercise T1.2 in Chapter 1, provides an alternative definition that is used throughout the text.

Table A.1: Some useful trigonometric identities.

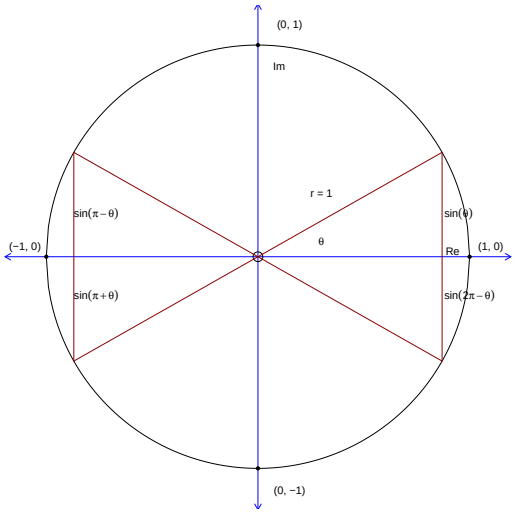
Double-angle Identities		
$\sin(2\theta) = 2\sin\theta\cos\theta = 2\tan\theta/(1+\tan^2\theta)$		$\cos(2\theta) = \cos^2\theta - \sin^2\theta = 1 - 2\sin^2\theta$
$\tan(2\theta) = 2\tan\theta/(1-\tan^2\theta)$		$\cot(2\theta) = (\cot^2\theta - 1)/(2\cot\theta)$
Half-angle Identities		
$\sin(\theta/2) = \pm\sqrt{(1-\cos\theta)/2}$		$\cos(\theta/2) = \pm\sqrt{(1+\cos\theta)/2}$
$\tan(\theta/2) = \pm\sqrt{(1-\cos\theta)/(1+\cos\theta)} = \sin\theta/(1+\cos\theta) = (1-\cos\theta)/\sin\theta$		
Product-to-Sum Identities		
$2\cos\alpha\cos\beta = \cos(\alpha+\beta) + \cos(\alpha-\beta)$		$2\sin\alpha\sin\beta = \cos(\alpha-\beta) - \cos(\alpha+\beta)$
$2\sin\alpha\cos\beta = \sin(\alpha+\beta) + \sin(\alpha+\beta)$		$2\cos\alpha\sin\beta = \sin(\alpha+\beta) - \sin(\alpha-\beta)$
Sum-to-Product Identities		
$\sin\alpha + \sin\beta = 2\sin[(\alpha+\beta)/2]\cos[(\alpha-\beta)/2]$		$\cos\alpha + \cos\beta = 2\cos[(\alpha+\beta)/2]\cos[(\alpha-\beta)/2]$
$\sin\alpha - \sin\beta = 2\sin[(\alpha-\beta)/2]\cos[(\alpha+\beta)/2]$		$\cos\alpha - \cos\beta = -2\sin[(\alpha+\beta)/2]\sin[(\alpha-\beta)/2]$
$\tan(\alpha \pm \beta) = (\tan\alpha \pm \tan\beta)/(1 \mp \tan\alpha\tan\beta)$		
Other Useful Identities		
$2\sin^2\theta = 1 - \cos(2\theta)$	$2\cos^2\theta = 1 + \cos(2\theta)$	$\sin^2\theta + \cos^2\theta = 1$ (Pythagorean Identity)

When the measure of θ is restricted to be between 0 and $\pi/2$, some interesting relationships between the previously defined trigonometric functions are the following.

$$\begin{aligned} \sin\theta &= \cos\left(\frac{\pi}{2} - \theta\right) \\ \cos\theta &= \sin\left(\frac{\pi}{2} - \theta\right) \\ \tan\theta &= \cot\left(\frac{\pi}{2} - \theta\right) \end{aligned}$$

Identities such as these, that interrelate trigonometric functions are referred to as **trigonometric identities**. Table A.1 contains a set of trigonometric identities that are useful in the some of the exercises and proofs in this text, especially those that relate to the spectral analysis of time series. There are many other identities. Common identities that are omitted from this table can, for the most part, be easily derived using those identities that do appear in the table. In the last row of the table is the "Pythagorean Identity," which is easily proven using the Pythagorean Theorem.

Figure A.1 illustrate that the trigonometric functions can be translated onto the complex mathematical plane, where the horizontal axis represents the space of real numbers and the vertical axis the space of imaginary numbers. A famous result, known as Euler's formula,



state that for any real number θ,

$$e^{i\theta} = \cos\theta + i\sin\theta, \tag{A.7}$$

where $i = \sqrt{-1}$ is the **imaginary number**. For an imaginary number $z = a + bi$, let $\Re(z) = a$ be the real part of z and $\Im(z) = b$ be the imaginary part of z. Then, for any real number θ, a little algebra will show that the sine and cosine functions may be written

$$\cos\theta = \Re(e^{i\theta}) = \frac{e^{i\theta} + e^{-i\theta}}{2} \tag{A.8}$$

$$\sin\theta = \Im(e^{i\theta}) = \frac{e^{i\theta} - e^{-i\theta}}{2i}. \tag{A.9}$$

A.2 Fourier Series and Spectral Representations

The ideas of frequency domain analysis of data are an important part of time series analysis. We have attempted to motivate these ideas intuitively in the main part of this book. In this appendix, we describe in more detail the mathematics behind the ideas. For more details, the reader is encouraged to consult Chapter 4 of Priestley (1981) or Chapter 7 of Anderson (1994). We recall that the basic idea of the frequency domain analysis of time series is to express various quantities as the "sum" of orthogonal sinusoids.

A.2.1 Fourier Series for Periodic Functions

Suppose that f is a function defined on a finite interval $[a, b]$. By this we also include functions that are periodic or period $b-a$. Throughout this book we have taken our interval to be [0, 1]. If it is not, we can construct the function $g(y) = f[(x-a)/(b-a)]$ and apply the results to the function g. The interval [0, 1] is natural for spectral densities as then frequency can be thought of as cycles per unit of time. Thus for monthly data, frequency 1/12 refers to possible cycles of length 12 months. The interval traditionally considered in mathematics is $[-\pi, \pi]$, which means for time series that one has to perform mental arithmetic involving π in order to convert frequency to a physically meaningful quantity. Using the interval [0, 1] also simplifies many of the formulas in the subject.

Our aim is to decompose f into the sum of sinusoids that are independent in some sense. Thus we begin by trying to approximate f by a sum of sinusoids of periods $1, 1/2, \ldots, 1/n$ for an arbitrary positive integer n.

Theorem A.2.1 Fourier Series Approximation.
Let f be a square integrable function defined on [0, 1]; that is $\int_0^1 f^2(x)\,dx < \infty$. Let

$$g_j(x) = a_j \cos(2\pi jx) + b_j \sin(2\pi jx) = C_j \cos(2\pi jx - \phi_j),$$

be a sinusoid of period $1/j$. Let

$$f_n(x) = \frac{a_0}{2} + \sum_{j=1}^{n} g_j(x),$$

where a_0 is some constant. Then the values of $a_0, a_1, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ that minimize

$$\int_0^1 \left| f(x) - \frac{a_0}{2} - \sum_{j=1}^{n} g_j(x) \right|^2 dx$$

are given by

$$a_j = 2\int_0^1 \cos(2\pi jx)\, f(x)\,dx, \quad j = 0, 1, \ldots, n \tag{A.10}$$

$$b_j = 2\int_0^1 \sin(2\pi jx)\, f(x)\,dx, \quad j = 1, \ldots, n. \tag{A.11}$$

The theorem yields that the sum of sinusoids that is closest to f in the integrated square error sense (which is analogous to the sum of square errors in regression analysis) has the coefficients given in equations (A.10) and (A.11). Note that the function f need not be square integrable in order to calculate the coefficients. In fact, f needs only to be absolutely integrable; that is $\int_0^1 |f(x)|\,dx < \infty$, since (for example)

$$\int_0^1 |f(x)\cos(2\pi jx)|\,dx \le \int_0^1 |f(x)|\,|\cos(2\pi jx)|\,dx \le \int_0^1 |f(x)|\,dx < \infty,$$

since $|\cos(2\pi jx)| \le 1$. Note that a square integrable function is absolutely integrable, but not necessarily vice versa. Consider, for example, $f(x) = 1/\sqrt{x}$.

Definition A.2.1 Fourier Coefficients & Partial Sum Fourier Approximant.
Let f be an absolutely integrable function defined on $[0, 1]$*, and let*

$$\begin{aligned} a_j &= 2\int_0^1 \cos(2\pi jx)\, f(x)\, dx, \quad j = 0, 1, \ldots, \\ b_j &= 2\int_0^1 \sin(2\pi jx)\, f(x)\, dx, \quad j = 1, 2, \ldots, \end{aligned}$$

and

$$f_n(x) = \frac{a_0}{2} + \sum_{j=1}^{n} [a_j \cos(2\pi jx) + b_j \sin(2\pi jx)]\,.$$

Then the a_j and b_j are called the **cosine** *and* **sine Fourier coefficients of f**, *while f_n is called the* **n^{th} partial sum Fourier approximant to f**.

In general, we will have that $f \neq f_n$. To accomplish harmonic analysis of f then, we hope that $f_n \to f$ in some sense, and also that there is some sense in which the sinusoids $g_1, g_2, \ldots$ are orthogonal. In the next theorem, we present results showing that in many cases, we do have the convergence and orthogonality.

Theorem A.2.2 Properties of Fourier Series.
Let f be an absolutely integrable function defined on the interval $[0, 1]$ *and let $f_1, f_2, \ldots$ be it approximating Fourier sums. Then*

(a) The sinusoids $f_j(x) = a_j \cos(2\pi jx) + b_j \sin(2\pi jx)$ are orthogonal in the sense that

$$\int_0^1 g_j(x) g_k(x)\, dx = \begin{cases} 0, & j \neq k \\ a_j^2 + b_j^2, & j = k. \end{cases}$$

(b) We have the following results about pointwise convergence of $f_n(x)$.

(1) At a point x where f is continuous on the left and right derivatives of f exist,

$$\lim_{n\to\infty} f_n(x) = f(x).$$

(2) At a point x where f is discontinuous but the left and right derivatives of f exist,

$$\lim_{n\to\infty} f_n(x) = \frac{f(x^+) + f(x^-)}{2},$$

where $f(x^+)$ and $f(x^-)$ denote the right and left limits of f at the point x; that is, the approximating Fourier sums converge to the average of the two values of f at the discontinuity.

(3) If f is continuous and has square integrable derivative, then $f_n \to f$ absolutely and uniformly.

(c) If f is square integrable, then

$$\lim_{n\to\infty} \int_0^1 |f(x) - f_n(x)|^2 \, dx = 0,$$

while

$$2\int_0^1 f^2(x)\, dx = \frac{a_0^2}{2} + \sum_{k=1}^{\infty} \left(a_k^2 + b_k^2\right).$$

Implications: This theorem gives us our harmonic analysis for two types of functions – absolutely integrable (part (b)) and square integrable (part (c)). For square integrable functions $f_n \to f$ in the integrated squared error sense, which need not mean that $f_n(x)$ converges pointwise to $f(x)$. Note that the second equation in part (c) (which is called Perceval's identity) is analogous to a decomposition of "total sum of squares" in analysis of variance, where the integral plays the role of the sum, into a sum of squares of amplitudes of the sinusoids for periods $1, 1/2, 1/3, \ldots$. For absolutely integrable functions, part (b1) shows that under general conditions, $f_n(x)$ does converge to $f(x)$ pointwise, while part (b2) explains the behavior of the Fourier approximations for the ideal bandpass filter in Section 4.3. Thus at the two points of discontinuity in the transfer function, $f_n(x)$ converges to the average of 0 and 1; that is to 0.5. Finally, part (a) shows that as long as the Fourier coefficients exist, then the sinusoids are orthogonal where now the inner product is defined as an integral rather than as a sum as it was in Section 2.2.3.

The Complex Form of Fourier Series

Recall Euler's equation given in equation (A.7)

$$e^{i\theta} = \cos\theta + i\sin\theta,$$

along with equations (A.8) and (A.9). Applying these results to the previous discussion provides a complex form of Fourier series. Specifically, write

$$\cos(2\pi jx) = \frac{e^{2\pi ijx} + e^{-2\pi ijx}}{2} \quad \text{and} \quad \sin(2\pi jx) = \frac{(-i)\left(e^{2\pi ijx} - e^{-2\pi ijx}\right)}{2},$$

which gives

$$f_n(x) = \frac{a_0}{2} + \sum_{j=1}^{n} [a_j \cos(2\pi jx) + b_j \sin(2\pi jx)] = \sum_{j=-n}^{n} r_j e^{-2\pi ijx},$$

where

$$r_j = \frac{a_j + ib_j}{2} \quad \text{and} \quad r_{-j} = \bar{r}_k = \frac{a_j - ib_k}{2}, \quad j = 0, 1, \ldots.$$

Note that

$$r_j = \int_0^1 f(x) e^{2\pi ijx}\, dx.$$

Thus we have that f and the rs are Fourier pairs and

$$f(x) = \sum_{j=-\infty}^{\infty} r_j e^{-2\pi ijx} \quad \text{and} \quad r_j = \int_0^1 f(x) e^{2\pi ijx}\, dx.$$

The r_j are called the complex Fourier coefficients of f.

A.2.2 Spectral Representation of R

In Appendix A.2.1, we were concerned with whether a function could be represented as a sum of sinusoids. In time series, we have the converse problem; tat is, we begin with the autocovariance sequence and ask whether there is a function having the Rs as its Fourier coefficients. The simplest, although not the most general, answer to this question is given in the following theorem.

Theorem A.2.3 Spectral Representation of R.
If $\{R(v) : v \in \mathcal{Z}\}$ is a positive definite sequence of numbers such that $R(v) = R(-v)$ and $\sum_{v=-\infty}^{\infty} |R(v)| < \infty$, then there exists a continuous function f such that

(a) $f(\omega) = \sum_{v=-\infty}^{\infty} R(v) e^{-2\pi iv\omega}$ *for* $\omega \in [0,\, 1]$.

(b) $R(v) = \int_0^1 f(\omega) e^{2\pi iv\omega}\, d\omega$ *for* $v \in \mathcal{Z}$.

(c) $f(\omega) = f(1 - \omega)$.

This theorem gives a harmonic analysis of an absolutely summable autocovariance function in terms of the spectral density function f. Unfortunately, the covariance function is not always absolutely summable, as illustrated by a harmonic process in Section 5.2. We can still obtain a harmonic analysis of a non-absolutely summable R as an integral, but the integral is of a special type called a Lebesgue-Stilts integral. In the spectral representation of R in Theorem 4.2 we wrote

$$R_v = \int_0^1 \cos(2\pi v\omega)\, dF(\omega),$$

where the integral is of the Lebesgue-Stieltjes type. Instead of defining this type of integral, which is rather complicated, we describe in the next theorem a way to calculate and interpret it in the special cases with which we are concerned.

Theorem A.2.4 Lebesgue-Stieltjes Integral.
Let g and F be bounded functions defined on $[0,\, 1]$ and for a positive integer m, consider partitioning $[0,\, 1]$ into 2^m intervals

$$\left[0,\, \frac{1}{2^m}\right], \quad \left[\frac{1}{2^m},\, \frac{2}{2^m}\right], \quad \ldots, \left[\frac{2^m - 1}{2^m},\, 1\right].$$

Thus if we let $\omega_{m,j} = j/2^m$, *for* $j = 0, 1, \ldots, 2^m$, *the* k^{th} *interval is from* $\omega_{m,k-1}$ *to* $\omega_{m,k}$, *for* $k = 1, \ldots 2^m$. *Now let*

$$I_m = \sum_{k=1}^{2^m} g\left(\omega'_{m,k}\right) \left[F(\omega_{m,k}) - F(\omega_{m,k-1})\right],$$

where $\omega'_{m,k} = (\omega_{m,k-1} + \omega_{m,k})/2$. *Then if* g *is Lebesgue-Stieltjes integrable with respect to* F, *and* g *is continuous (and therefore also Riemann-Stieltjes integrable), we have that* $\lim_{m\to\infty} I_m$ *exists, is finite, and*

$$\int_0^1 g(x)\, dF(x) = \lim_{m\to\infty} I_m.$$

From this result we have that

$$R_v = \int_0^1 \cos(2\pi v\omega)\ dF(\omega) = \lim_{m\to\infty} \sum_{k=1}^{2^m} \cos\left(2\pi v\omega'_{m,k}\right) \left[F(\omega_{m,k}) - F(\omega_{m,k-1})\right],$$

which means that the autocovariance sequence R can be regarded as the sum of a large number of cosines with amplitudes given by "increments" of the spectral distribution function F. Thus, if there is a large increase in F between frequencies $\omega_{m,k-1}$ and $\omega_{m,k}$, the sinusoid having frequency $\omega'_{m,k}$ will be prominent in the decomposition of R. In fact, if the only large increase in F is in the interval containing $\omega'_{m,k}$, then R will appear almost exactly sinusoidal. The logical limit of this idea is the autocovariance of the harmonic process.

The Harmonic Process

Consider the harmonic process having a single frequency component of $\omega_0 \in (0,\, 0.5)$; that is, $R_v = \sigma^2 \cos(2\pi v\omega_0)$. Define

$$F(\omega) = \begin{cases} 0, & 0 \le \omega < \omega_0 \\ \dfrac{\sigma^2}{2}, & \omega_0 \le \omega < 1 - \omega_0 \\ \sigma^2 & 1 - \omega_0 \le \omega \le 1, \end{cases}$$

and let ω_m be the center of the interval in the partition of $[0,\, 1]$ at the m^{th} step of the limiting process. If ω_0 is on the boundary between two intervals, we choose ω_m to be the center of the one containing the jump in F. Thus not that $1 - \omega_m$ is in the center of the interval containing $1 - \omega_0$ and we have

$$\begin{aligned} & \lim_{m\to\infty} \sum_{k=1}^{2^m} \cos\left(w\pi v\omega'_{m,k}\right) \left[F\left(\omega_{m,k}\right) - F\left(\omega_{m,k-1}\right)\right] \\ = & \lim_{m\to\infty} \left[\frac{\sigma^2}{2} \cos\left\{2\pi v\omega_m\right\} + \frac{\sigma^2}{2} \cos\left\{2\pi v(1-\omega_m)\right\}\right] \\ = & \sigma^2 \cos\left(2\pi v\omega_0\right), \end{aligned}$$

since $\omega_m \to \omega_0$ and $\cos(2\pi v\omega) = \cos\{2\pi v(1-\omega)\}$. Thus this definition of the spectral distribution function does indeed lead to the correct autocovariance function. We note that in the general theory of representing sequences, the function F is unique, and thus our definition of F is the only one that represents R.

In general, if F is absolutely continuous, we can see that the Lebesgue-Stieltjes integral reduces to the usual Riemann integral since then

$$\begin{aligned}
\int_0^1 \cos(w\pi v\omega)\, dF(\omega) &= \lim_{m\to\infty} \frac{1}{2^m} \sum_{k=1}^{2^m} \cos\left(2\pi v\omega'_{m,k}\right) \left[\frac{F(\omega_{m,k}) - F(\omega_{m,k-1})}{1/2^m}\right] \\
&\doteq \lim_{m\to\infty} \frac{1}{2^m} \sum_{k=1}^{2^m} \cos\left(2\pi v\omega'_{m,k}\right) \frac{dF\left(\omega'_{m,k}\right)}{d\omega} \\
&= \int_0^1 \cos(2\pi v\omega)\, f(\omega)\, d\omega
\end{aligned}$$

by the definition of the Riemann integral.

A.2.3 Spectral Representation of X

It is possible to obtain a harmonic analysis of the time series X itself (see part (4) of Theorem 4.2). The representation is different than that for R in three ways. First, the representation of R is in terms of cosines only. This is because R is an even function; that is $R(v) = R(-v)$. Second, the increments $F(\omega_k) - F(\omega_j)$ are replaced by increments of two uncorrelated stochastic processes (more on this below). Finally, since the representation will be as a limit of sums of linear combinations of random variables, the usual limit cannot be used, rather the limit is a limit in mean square (see Definition e in Appendix A.6).

A stochastic process $\{C(\omega), \omega \in [0, 1]\}$ is an indexed collection of random variables. The random variable $C(\omega_2) - C(\omega_1)$ for $0 \le \omega_1 < \omega_2 \le 1$ is called an **increment** of C. If increments defined on non-overlapping intervals are uncorrelated, we say that C has **uncorrelated increments**.

Thus the spectral representation of X means essentially that

$$\begin{aligned}
X_t = \text{l.i.m.} \sum_{k=1}^{2^m} \Big[&\cos\left(2\pi v\omega'_{m,k}\right) \{C(\omega_{m,k}) - C(\omega_{m,k-1})\} \\
&+ \sin\left(2\pi v\omega'_{m,k}\right) \{S(\omega_{m,k}) - S(\omega_{m,k-1})\}\Big]
\end{aligned}$$

where l.i.m. denotes limit in mean square. This again allows us to think of a particular infinitely long realization from X as the sum of sinusoids whose amplitudes have been chosen according to some random mechanism, determined by the processes C and S, whose properties (such as variance) are described by the spectral distribution function F.

A.3 The Fast Fourier Transform

Many of the calculations in time series analysis require the calculation of the discrete Fourier transform of a set of numbers (see Section 2.2.3 for example). In this section we describe the fast Fourier transform (FFT). The basic ideas were developed and popularized by Cooley & Tukey (1965) and Gentleman & Sande (1966). The FFT is a classic example of the use of recursion to greatly reduce the number of operations required to solve a computational problem. Although there are a wide variety of algorithms known as FFTs, almost all of them have two essential elements: (1) a recursion that allows one to find the DFT of n points as a simple combination of DFTs for subsets of the points, and (2) a reordering of the original data so as to minimize the number of complex exponentials that need to be calculated and/or stored while the recursion is being performed.

A.3.1 The FFT Recursion

Let $W_n = e^{2\pi i/n}$. Then we can write the DFT of $X_1, \ldots, X_n$ as

$$Z_k = \sum_{t=1}^{n} X_t W_n^{(t-1)(k-1)}, \quad k = 1, \ldots, n.$$

Theorem A.3.1 The FFT Recursion.
Suppose n can be factored into $n = rs$. Define the $r \times s$ matrix $\boldsymbol{V}$ to have j^{th} column

$$X_j, X_{s+j}, \ldots, X_{(r-1)s+j}, \quad j = 1, \ldots, s.$$

Then

$$Z_k = \sum_{\ell=1}^{s} W_n^{(k-1)(\ell-1)} Z_{\ell, mod(k-1,r)+1},$$

where $Z_{\ell,1}, \ldots, Z_{\ell,r}$ is the DFT of the ℓ^{th} column of $\boldsymbol{V}$.

Proof: We can get Z_k by adding over the elements of $\boldsymbol{V}$:

$$\begin{aligned} Z_k &= \sum_{\ell=1}^{s} \sum_{m=1}^{r} V_{m\,\ell} W_n^{(k-1)[(m-1)s+(\ell-1)]} \\ &= \sum_{\ell=1}^{s} W_n^{(k-1)(\ell-1)} \sum_{m=1}^{r} V_{m\,\ell} W_n^{(k-1)(m-1)s} \end{aligned}$$

since $V_{m\,\ell} = X_{(m-1)s+\ell}$. But $s/n = 1/r$ which gives $W_n^s = W_r$. Furthermore

$$k - 1 = \left[\frac{k-1}{r}\right] r + (k-1) - \left[\frac{k-1}{r}\right] r = \left[\frac{k-1}{r}\right] r + \mathrm{mod}(k-1, r),$$

and thus

$$W_n^{(k-1)(m-1)s} = W_r^{(m-1)\left\{\left[\frac{k-1}{r}\right] r + \mathrm{mod}(k-1,r)\right\}} = W_r^{(m-1)\mathrm{mod}(k-1,r)},$$

since $W_r^{\left[\frac{k-1}{r}\right]r} = 1$ and $e^{2\pi i} = 1$. Thus

$$\begin{aligned} Z_k &= \sum_{\ell=1}^{s} W_n^{(k-1)(\ell-1)} \sum_{m=1}^{r} V_{m\,\ell} W_r^{(m-1)\bmod(k-1,r)} \\ &= \sum_{\ell=1}^{s} W_n^{(k-1)(\ell-1)} Z_{\ell,\bmod(k-1,r)+1}. \end{aligned}$$

Thus it takes $sr^2 = nr$ operations to calculate DFTs of all s columns of $\boldsymbol{V}$ and then ns operations to calculate all n Zs, making a total of $nr + ns = n(r + s)$ operations instead of n^2. Further, if r is itself factorable, we can use the same idea to more rapidly find the DFTs of the columns of $\boldsymbol{V}$. This process can be continued until all of the prime factors of n are exhausted, giving a total of $n(p_1 + \cdots + p_K)$ operations, where $p_1, \ldots, p_K$ are the prime factors of n.

A.3.2 Reordering Data

The recursions given in Theorem A.3.1 can be further improved if we can order the calculations in such a way that all of the one using a certain complex exponential can be done at the same time. To illustrate this, consider the case of $n = 8$ and $s = 2$:

$$\begin{aligned} Z_1 &= Z_{1,1} + W_8^0 Z_{2,1} \\ Z_2 &= Z_{1,2} + W_8^1 Z_{2,2} \\ Z_3 &= Z_{1,3} + W_8^2 Z_{2,3} \\ Z_4 &= Z_{1,4} + W_8^3 Z_{2,4} \\ Z_5 &= Z_{1,1} + W_8^4 Z_{2,1} \\ Z_6 &= Z_{1,2} + W_8^5 Z_{2,2} \\ Z_7 &= Z_{1,3} + W_8^6 Z_{2,3} \\ Z_8 &= Z_{1,4} + W_8^7 Z_{2,4}, \end{aligned}$$

where $Z_{1,\cdot}$ is the DFT of the odd numbered Xs and $Z_{2,\cdot}$ is the DFT of the even numbered Xs. Thus

$$\begin{aligned} Z_{1,1} &= Z_{1,1}^{(1)} + W_4^0 Z_{1,1}^{(2)} & \qquad Z_{2,1} &= Z_{2,1}^{(1)} + W_4^0 Z_{2,1}^{(2)} \\ Z_{1,2} &= Z_{1,2}^{(2)} + W_4^1 Z_{1,2}^{(2)} & \qquad Z_{2,2} &= Z_{2,2}^{(1)} + W_4^1 Z_{2,2}^{(2)} \\ Z_{1,3} &= Z_{1,1}^{(1)} + W_4^2 Z_{1,1}^{(2)} & \qquad Z_{2,3} &= Z_{2,1}^{(1)} + W_4^2 Z_{2,1}^{(2)} \\ Z_{1,4} &= Z_{1,2}^{(2)} + W_4^3 Z_{1,2}^{(2)} & \qquad Z_{2,4} &= Z_{2,2}^{(1)} + W_4^3 Z_{2,2}^{(2)} \end{aligned}$$

where

$$\begin{aligned} Z_{1,1}^{(1)} &= X_1 + W_2^0 X_5 \\ Z_{1,2}^{(1)} &= X_1 + W_2^1 X_5 \\ Z_{1,1}^{(2)} &= X_3 + W_2^0 X_7 \end{aligned}$$

$$
\begin{aligned}
Z_{1,2}^{(2)} &= X_3 + W_2^1 X_7 \\
Z_{2,1}^{(1)} &= X_2 + W_2^0 X_6 \\
Z_{2,2}^{(1)} &= X_2 + W_2^1 X_6 \\
Z_{2,1}^{(2)} &= X_4 + W_2^0 X_8 \\
Z_{2,2}^{(2)} &= X_4 + W_2^1 X_8.
\end{aligned}
$$

Thus if we start by permuting the Xs, we can summarize the calculations by

$$
\begin{bmatrix} X_1 \\ X_5 \\ X_3 \\ X_7 \\ X_2 \\ X_6 \\ X_4 \\ X_8 \end{bmatrix}
\longrightarrow
\begin{bmatrix} Z_{1,1}^{(1)} \\ Z_{1,1}^{(2)} \\ Z_{2,1}^{(1)} \\ Z_{2,1}^{(2)} \\ Z_{1,2}^{(1)} \\ Z_{1,2}^{(2)} \\ Z_{2,2}^{(1)} \\ Z_{2,2}^{(2)} \end{bmatrix}
\longrightarrow
\begin{bmatrix} Z_{1,1} \\ Z_{2,1} \\ Z_{1,2} \\ Z_{2,2} \\ Z_{1,3} \\ Z_{2,3} \\ Z_{1,4} \\ Z_{2,4} \end{bmatrix}
\longrightarrow
\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \\ Z_8 \end{bmatrix},
$$

where the first arrow means first combine all four consecutive pairs using W_2^0 and then again using W_2^1. The second arrow means combine the first two pairs within W_4^0, the next two using W_4^1, the next two using W_4^2, and the last two using W_4^3. In the last step we combine the first pair using W_8^0, the next pair using W_8^1, and so on. Thus at each step we are fixing the complex exponential and then doing all combinations using that value.

To determine the index k of X_j in the permuted array, one need only write down the binary representation of $j-1$, reverse the bits, and then add one. For example, X_5 is mapped to position 2 since $5 = 1 = 4$, and the binary representation of 4 is $100 \to 001 \to 010 = 2$. Note that if X_j gets mapped to position k, then X_k gets mapped to position j.

There are many variations on this theme and further simplifications that can be made. For some of the functions in *Timeslab in R*, we use the algorithm described by Singleton (1969) which is in some ways still the standard by which newer methods are judged. We chose this algorithm because it performs the transform in place and will work for a wide variety of different prime factors, although as we saw in Section 2.2.4, it does require that n not have any large prime factors. Other functions in *Timeslab in R* use an FFT that is the default in *R*. This version is a modification of the Singleton (1969) FFT that was presented in Singleton (1979).

A.4 Matrix Operations

In this section we consider three operations on matrices: (1) the modified Cholesky decomposition of a positive definite matrix, (2) the Gram-Schmidt decomposition of a nonsingular $n \times m$ matrix, and (3) the matrix sweep operator. Each of these is useful as both a computational and theoretical tool.

A.4.1 The Modified Cholesky Decomposition

An $n \times n$ matrix $\boldsymbol{V}$ is said to be positive definite or positive semidefinite if for any nonzero n-dimensional vector $\boldsymbol{h}$, we have $\boldsymbol{h}^T\boldsymbol{V}\boldsymbol{h} > 0$ of $\boldsymbol{h}^T\boldsymbol{V}\boldsymbol{h} \geq 0$ respectively. In statistics, it is often necessary to find what is called the square root of a positive definite matrix.

Definition A.4.1 Square Root of a Matrix.
The **square root of a symmetric, positive definite** $n \times n$ **matrix** $\boldsymbol{V}$ *is an* $n \times n$ *matrix* $\boldsymbol{A}$ *satisfying* $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^T$. *We denote such a matrix by* $\boldsymbol{V}^{1/2}$. *The inverse square root of* $\boldsymbol{V}$ *is the inverse of* $\boldsymbol{V}^{1/2}$ *and is denoted* $\boldsymbol{V}^{-1/2}$. *The transposes of* $\boldsymbol{V}^{-1}$ *and* $\boldsymbol{V}^{-1/2}$ *are denoted* $\boldsymbol{V}^{-T}$ *and* $\boldsymbol{V}^{-T/2}$ *respectively.*

The next theorem shows that there exists a unique matrix square root for any positive definite matrix and also shows how to find it. Note that a triangular matrix is called **unit** if it has ones on the main diagonal.

Theorem A.4.1 Modified Cholesky Decomposition.
Let $\boldsymbol{V}$ *be a symmetric* $n \times n$ *matrix. Then*

(a) $\boldsymbol{V}$ *is positive definite if and only if there exists a unique unit lower triangular* $n \times n$ *matrix* $\boldsymbol{L}$ *and a unique diagonal* $n \times n$ *matrix* $\boldsymbol{D}$ *having positive diagonal elements, such that*

$$\boldsymbol{V} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T.$$

(b) This factorization is called the modified Cholesky decomposition (MCD) of $\boldsymbol{V}$, *and if the decomposition exists we can calculate the elements* $\boldsymbol{L}$ *and* $\boldsymbol{D}$ *one row at a time by* $D_{11} = V_{11}$ *and for* $i = 2, \ldots, n$

$$\begin{aligned} L_{ij} &= \frac{V_{ij} - \sum_{\ell=1}^{j-1} L_{i\ell} D_{\ell\ell} L_{j\ell}}{D_{jj}}, \quad j = 1, \ldots, i-1 \\ D_{ii} &= V_{ii} = \sum_{\ell=1}^{i-1} D_{\ell\ell} L_{i\ell}^2. \end{aligned}$$

(c) The MCD of $\boldsymbol{V}$ *is nested; that is, if* $\boldsymbol{V}_k, \boldsymbol{L}_k$, *and* $\boldsymbol{D}_k$ *represent the upper left-hand* $k \times k$ *parts of* $\boldsymbol{V}, \boldsymbol{L}$ *and* $\boldsymbol{D}$, *respectively, then*

$$\boldsymbol{V}_k = \boldsymbol{L}_k \boldsymbol{D}_k \boldsymbol{L}_k^T, \quad k = 1, \ldots, n,$$

and thus for and k *greater than or equal to* i *and* j, *the* $(i,j)^{th}$ *elements of* $\boldsymbol{V}_k, \boldsymbol{L}_k$ *and* $\boldsymbol{D}_k$ *can be denoted by* V_{ij}, L_{ij}, *and* D_{ij}, *respectively.*

(d) The unique square root of a positive definite matrix $\boldsymbol{V}$ *is given by* $\boldsymbol{V}^{1/2} = \boldsymbol{L}\boldsymbol{D}^{1/2}$ *where* $\boldsymbol{D}^{1/2} = Diag(D_{11}^{1/2}, \ldots, D_{nn}^{1/2})$.

(e) The inverse square root of a positive definite matrix $\boldsymbol{V}$ *is given by* $\boldsymbol{V}^{-1/2} = \boldsymbol{D}^{-1/2}\boldsymbol{L}^{-1}$ *where* $\boldsymbol{D}^{-1/2} = Diag(D_{11}^{-1/2}, \ldots, D_{nn}^{-1/2})$.

The `mchol` Function

If `A` is a symmetric $n \times n$ matrix ($\boldsymbol{A}$), then the function

```
mchol(A)
```

will attempt to find the modified Cholesky factors $\boldsymbol{L}$ and and $\boldsymbol{D}$ of $\boldsymbol{A}$. It returns a list containing three objects: `error`, `L`, and `D`. If the algorithm judges that $\boldsymbol{A}$ is not positive definite, that is, if a diagonal element $\boldsymbol{D}$ is found that is less than 10^{-25}, the function will return an error value of `error = 1` with null values for `L` and `D`. If $\boldsymbol{A}$ is judged to be positive definite, then `mchol` will return `error = 0`, and will return the factors $\boldsymbol{L}$ and $\boldsymbol{D}$ in the matrices `L` and `D`. This in addition to finding the decomposition in possible, `mchol` provides an easy check for positive definiteness of a matrix.

A.4.2 The Gram-Schmidt Decomposition

An important operation in many areas of mathematics and statistics is to find what is called an orthogonal basis for a matrix; that is, from one set of vectors, find a new set by some linear transformation such that the inner product of any two different new vectors is zero. We will use the Gram-Schmidt decomposition (see Clayton (1971)) to accomplish this aim.

Theorem A.4.2 Gram-Schmidt Decomposition.
If $\boldsymbol{X}$ is an $n \times p$ matrix having full rank p, then there exists an $n \times p$ orthogonal matrix $\boldsymbol{Q}$ (that is, $\boldsymbol{Q}^T\boldsymbol{Q}$ is diagonal) and a unit upper triangular $p \times p$ matrix $\boldsymbol{R}$ such that

$$\boldsymbol{X} = \boldsymbol{QR}.$$

This decomposition of $\boldsymbol{X}$ is called its **Gram-Schmidt decomposition** *(GSD).*

Note that $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{QR} = \boldsymbol{R}^T\boldsymbol{VR}$ where $\boldsymbol{R}^T$ is unit lower triangular and $\boldsymbol{V}$ is diagonal with positive diagonal elements. Since the modified Cholesky decomposition $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{LDL}^T$ of $\boldsymbol{X}^T\boldsymbol{X}$ is unique, we have that $\boldsymbol{R}^T = \boldsymbol{L}$ and $\boldsymbol{V} = \boldsymbol{D}$.

The `gram.schmidt` Function

If the matrix `A` is $n \times p$, then the command

```
gram.schmidt(A)
```

will attempt to find the factors returns a list containing three objects: `error`, `Q`, and `R`. If there is an error, then `error = 1`, and the values of `Q` and `R` will be null. If there is no error, then `error = 0`, and `Q` will contain the $\boldsymbol{Q}$ factor, and `R` the $\boldsymbol{R}$ factor of the GSD.

The function uses what is known as the "Gram-Schmidt decomposition algorithm". In other words, if we let $\boldsymbol{Q}_0 = \boldsymbol{X}$, and $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p = \boldsymbol{Q}$ be a sequence of $n \times p$ matrices, and

$\boldsymbol{q}_{ij}$ denote the j^{th} column of $\boldsymbol{Q}_i$, then `gram.schmidt` calculates for $i = 1, \ldots, p-1$

$$\begin{aligned} d_i &= \boldsymbol{q}_{i-1,i}^T \boldsymbol{q}_{i-1,i} \\ R_{ij} &= \frac{\boldsymbol{q}_{i-1,j}^T \boldsymbol{q}_{i-1,i}}{d_i}, \quad j = i+1, \ldots, p \\ \boldsymbol{q}_{ij} &= \boldsymbol{q}_{i-1,j} - R_{ij} \boldsymbol{q}_{i-1,i}, \quad j = i+1, \ldots, p. \end{aligned}$$

If $|d_i| < 10^{-25}$ for any i, then `gram.schmidt` concludes that $\boldsymbol{X}$ is singular, and an error will occur (`error = 1`).

A.4.3 The Sweep Operator

Many of the computational and theoretical results in regression, multivariate analysis, and time series can be expressed succinctly using what is called the sweep operator.

Definition A.4.2 Sweep Operator.
Let $\boldsymbol{A}$ be an $n \times n$ matrix. The process of sweeping $\boldsymbol{A}$ on its k^{th} diagonal element, denoted by $\boldsymbol{B} = SWEEP(k)\,\boldsymbol{A}$, is the process of forming the matrix $\boldsymbol{B}$ by (assuming $A_{kk} \neq 0$)

$$\begin{aligned} B_{kk} &= \frac{1}{A_{kk}} \\ B_{ik} &= -\frac{A_{ik}}{A_{kk}}, \quad i \neq k \quad (k^{th}\ column) \\ B_{kj} &= \frac{A_{kj}}{A_{kk}}, \quad j \neq k \quad (k^{th}\ row) \\ B_{ij} &= A_{ij} - \frac{A_{ik}A_{kj}}{A_{kk}}, \quad i \neq k,\ j \neq k. \end{aligned}$$

If $\boldsymbol{B}_1 = SWEEP(k_1)\,\boldsymbol{A}$, $\boldsymbol{B}_2 = SWEEP(k_2)\,\boldsymbol{B}_1, \ldots, \boldsymbol{B} = SWEEP(k_r)\,\boldsymbol{B}_{r-1}$, we write $\boldsymbol{B} = SWEEP(k_1, \ldots, k_r)\,\boldsymbol{A}$, and say that $\boldsymbol{B}$ is the result of sweeping $\boldsymbol{A}$ on its diagonals $k_1, \ldots, k_r$.

The sweep operator has many useful properties, as we see in the next theorem.

Theorem A.4.3 Properties of the Sweep Operator.
Let

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{D} & \boldsymbol{E} \end{bmatrix},$$

where $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$, and $\boldsymbol{E}$ are $r \times r, r \times s, s \times r, ands \times s$, respectively. Then

(a) $SWEEP(k_1, \ldots, k_p)\,\boldsymbol{A}$ can be found by sweeping the diagonals in any order, for example,

$$SWEEP(k_1,\, k_2)\,\boldsymbol{A} = SWEEP(k_2,\, k_1)\,\boldsymbol{A}.$$

(b) $SWEEP(k, k)\,\boldsymbol{A} = \boldsymbol{A}$; that is, sweeping a second time on a given diagonal undoes the effect of a previous sweeping on that diagonal.

(c) $SWEEP(1,\ldots,r)\,\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B}^{-1} & \boldsymbol{B}^{-1}\boldsymbol{C} \\ -\boldsymbol{D}\boldsymbol{B}^{-1} & \boldsymbol{E}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{C} \end{bmatrix}$ *if* $\boldsymbol{B}$ *is nonsingular.*

(d) $SWEEP(r+1,\ldots,r+s)\,\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B}-\boldsymbol{C}\boldsymbol{E}^{-1}\boldsymbol{D} & -\boldsymbol{C}\boldsymbol{E}^{-1} \\ \boldsymbol{E}^{-1}\boldsymbol{D} & \boldsymbol{E}^{-1} \end{bmatrix}$ *if* $\boldsymbol{E}$ *is nonsingular.*

(e) $SWEEP(1,\ldots,r+s)\,\boldsymbol{A} = \boldsymbol{A}^{-1}$ *and*

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{B}^{-1}+\boldsymbol{B}^{-1}\boldsymbol{C}(\boldsymbol{E}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}\boldsymbol{D}\boldsymbol{B}^{-1} & -\boldsymbol{B}^{-1}\boldsymbol{C}(\boldsymbol{E}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{C})^{-1} \\ -(\boldsymbol{E}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}\boldsymbol{D}\boldsymbol{B}^{-1} & (\boldsymbol{E}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{C})^{-1} \end{bmatrix}$$

if $\boldsymbol{B}, \boldsymbol{E}$, *and* $(\boldsymbol{E}-\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}$ *are nonsingular.*

The sweep operator can be used to solve many theoretical and computational problems. We consider two examples of its use. First, the basic quantities for the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ (where $\boldsymbol{X}$ is $n \times p$) can be found by applying property (c) (see Appendix A.5).

$$\text{SWEEP}\,(1,\ldots,p)\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{X} & \boldsymbol{y}^T\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\text{Var}\left(\hat{\boldsymbol{\beta}}\right) & \hat{\boldsymbol{\beta}} \\ -\hat{\boldsymbol{\beta}}^T & \text{RSS} \end{bmatrix}.$$

To illustrate the theoretical utility of sweep, notice for matrix $\boldsymbol{A}$ in Theorem A.4.3 that sweeping on all of its diagonals is the same as sweeping the matrix SWEEP $(r+1,\ldots,r+s)\,\boldsymbol{A}$ on diagonals $1,\ldots,r$. Thus the upper left-hand corner in part (e) is the inverse of the upper left-hand corner in part (d) by doing part (c). Doing a similar calculation with $\pm\boldsymbol{E}^{-1}$ replacing $\boldsymbol{E}$ in $\boldsymbol{A}$ gives the well-known matrix inversion lemma (see Rao (1973), p. 33, for example).

Theorem A.4.4 Matrix Inversion Lemma.
Let $\boldsymbol{B}$ *and* $\boldsymbol{E}$ *be nonsingular* $r \times r$ *and* $s \times s$ *matrices and* $\boldsymbol{C}$ *and* $\boldsymbol{D}$ *be* $r \times s$ *and* $s \times r$ *matrices. Then*

$$(\boldsymbol{B} \pm \boldsymbol{C}\boldsymbol{E}\boldsymbol{D})^{-1} = \boldsymbol{B}^{-1} \mp \boldsymbol{B}^{-1}\boldsymbol{C}(\boldsymbol{E}^{-1} \pm \boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}.$$

We will use this formula extensively when we consider recursive regression in Appendix A.5, that is, when we consider adjusting estimators for the addition or deletion of observations. It can also be used to motivate the Kalman Filter Algorithm discussed in Appendix A.6.

The `swp` Function

The `swp` function is called using

```
swp(A, k1 = 1, k2 = ncol(A))
```

where `A` is a matrix, `k1` is the index of the first diagonal element of `A` on which to begin sweeping, and `k2` is the index of the last diagonal. The default for `k1 = 1`, and for `k2` is the last column of `A`. The function returns a list containing the error indicator `error`, and a matrix `A`. If `error = 0`, then `A` is the result of sweeping the input matrix on diagonals `k1` through `k2` and `error`. If there is a diagonal element that is less than 10^{-25}, then `error = 1`, and the matrix `A` will be null.

A.5 Least Squares Regression Analysis

A.5.1 The General Linear Model

A.5.2 Results for the General Linear Model

A.5.3 The Case of Correlated Errors

A.5.4 Special Types of Regression Analysis

Orthogonal Regression

Stepwise Regression

Recursive Regression

An important topic in statistics is recursive estimation. In regression analysis, this means given the results $\hat{\boldsymbol{\beta}}, RSS$, and $\boldsymbol{V} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ of regressing the $m \times 1$ vector $\boldsymbol{y}$ on the $m \times p$ matrix $\boldsymbol{X}$, we seek the corresponding quantities after adding or deleting data $\boldsymbol{y}_1$, and $\boldsymbol{X}_1$ of length n. Thus we seek

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_N &= \left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{y} \pm \boldsymbol{X}_1^T\boldsymbol{y}_1\right) \\
RSS_N &= \left(\boldsymbol{y}^T\boldsymbol{y} \pm \boldsymbol{y}_1^T\boldsymbol{y}_1\right) - \hat{\boldsymbol{\beta}}_N^T\left(\boldsymbol{X}^T\boldsymbol{y} \pm \boldsymbol{X}_1^T\boldsymbol{y}_1\right) \\
\boldsymbol{V}_n &= \left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}
\end{aligned}$$

where the $\pm$ is $+$ for adding observations and $-$ for deleting. From the matrix inversion lemma (Theorem A.4.4) we can obtain recursive formulas for these quantities.

Theorem A.5.1 Recursive Regression Formulas.
The updated quantities $\hat{\boldsymbol{\beta}}_N, RSS_N$, and $\boldsymbol{V}_N$ defined above are given by

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_N &= \hat{\boldsymbol{\beta}} \pm \boldsymbol{A}\boldsymbol{X}_1^T\left(\boldsymbol{I}_n \pm \boldsymbol{H}_1\right)^{-1}\boldsymbol{e}_1 \\
&= \hat{\boldsymbol{\beta}} \pm \left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T\boldsymbol{e}_1 \\
RSS_N &= RSS \pm \boldsymbol{e}_1^T\left(\boldsymbol{I}_n \pm \boldsymbol{H}_1\right)^{-1}\boldsymbol{e}_1 \\
&= RSS \pm \boldsymbol{e}_1^T\left[\boldsymbol{I}_n \mp \boldsymbol{X}_1\left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T\right]\boldsymbol{e}_1 \\
\boldsymbol{V}_N &= \boldsymbol{A} \mp \boldsymbol{A}\boldsymbol{X}_1^T\left(\boldsymbol{I}_n \pm \boldsymbol{H}_1\right)^{-1}\boldsymbol{X}_1\boldsymbol{A} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1},
\end{aligned}$$

where

$$\boldsymbol{H}_1 = \boldsymbol{X}_1\boldsymbol{A}\boldsymbol{X}_1^T, \qquad \boldsymbol{A} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}, \qquad \boldsymbol{e}_1 = \boldsymbol{y}_1 - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}.$$

Proof: The basis of this theorem is the Matrix Inversion Lemma (Theorem A.4.4) and the facts below, easily verified by multiplication.

$$(\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1} = \boldsymbol{I}_n \mp (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1} \boldsymbol{H}_1 = \boldsymbol{I}_n \mp \boldsymbol{H}_1 (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}.$$

Thus

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_N &= \left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{y} \pm \boldsymbol{X}_1^T\boldsymbol{y}_1\right) \\
&= \left[\boldsymbol{A} \mp \boldsymbol{A}\boldsymbol{X}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1} \boldsymbol{X}_1\boldsymbol{A}\right]\left(\boldsymbol{X}^T\boldsymbol{y} \pm \boldsymbol{X}_1^T\boldsymbol{y}_1\right) \\
&= \hat{\boldsymbol{\beta}} \mp \boldsymbol{A}\boldsymbol{X}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1} \hat{\boldsymbol{y}}_1 \pm \boldsymbol{A}\boldsymbol{X}_1^T\boldsymbol{y}_1 - \boldsymbol{Z}\boldsymbol{X}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1} \boldsymbol{H}_1\boldsymbol{y}_1,
\end{aligned}$$

where $\hat{\boldsymbol{y}}_1 = \boldsymbol{X}_1\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}$, which gives

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}} &= \pm\boldsymbol{A}\boldsymbol{X}_1^T\left[\boldsymbol{I}_n \mp (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{H}_1\right]\boldsymbol{y}_1 \mp \boldsymbol{A}\boldsymbol{X}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\hat{\boldsymbol{y}}_1 \\
&= \pm\boldsymbol{A}\boldsymbol{X}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}(\boldsymbol{y}_1 - \hat{\boldsymbol{y}}_1),
\end{aligned}$$

which gives the first expression for $\hat{\boldsymbol{\beta}}_N$. To show the second, note that

$$\begin{aligned}
\left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T\boldsymbol{e}_1 &= \left[\boldsymbol{A} \mp \boldsymbol{A}\boldsymbol{X}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{X}_1\boldsymbol{A}\right]\boldsymbol{X}_1^T\boldsymbol{e}_1 \\
&= \boldsymbol{A}\boldsymbol{X}_1^T\left[\boldsymbol{I}_n \mp (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{H}_1\right]\boldsymbol{e}_1 \\
&= \boldsymbol{A}\boldsymbol{X}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{e}_1.
\end{aligned}$$

To find RSS_N, we have that $\pm\boldsymbol{H}_1\boldsymbol{y}_1 = (\boldsymbol{I}_n \pm \boldsymbol{H}_1)\boldsymbol{y}_1 - \boldsymbol{y}_1$, and so

$$\begin{aligned}
RSS_N &= \left(\boldsymbol{y}^T\boldsymbol{y} \pm \boldsymbol{y}_1^T\boldsymbol{y}_1\right) - \left[\hat{\boldsymbol{\beta}}^T \pm \boldsymbol{e}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{X}_1\boldsymbol{A}\right]\left(\boldsymbol{X}^T\boldsymbol{y} \pm \boldsymbol{X}_1^T\boldsymbol{y}_1\right) \\
&= RSS \pm \boldsymbol{y}_1^T\boldsymbol{y}_1 \mp \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_1^T\boldsymbol{y}_1 \mp \boldsymbol{e}_1 (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\left(\boldsymbol{X}_1\hat{\boldsymbol{\beta}} \pm \boldsymbol{H}_1\boldsymbol{y}_1\right) \\
&= RSS \pm \boldsymbol{y}_1^T\boldsymbol{y}_1 \mp \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_1^T\boldsymbol{y}_1 \mp \boldsymbol{e}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{y}_1^T \\
&\qquad \mp\boldsymbol{e}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\left[(\boldsymbol{I}_n \pm \boldsymbol{H}_1)\boldsymbol{y}_1 - \boldsymbol{y}_1\right] \\
&= RSS \pm \left(\boldsymbol{y}_1^T - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_1^T - \boldsymbol{e}_1^T\right)\boldsymbol{y}_1 \mp \boldsymbol{e}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)\left(\boldsymbol{y}_1^T - \boldsymbol{y}_1\right) \\
&= RSS \pm \boldsymbol{e}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{e}_1,
\end{aligned}$$

since $\boldsymbol{e}_1^T = \boldsymbol{y}_1^T - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_1^T$. To obtain the second expression for RSS_N, note that

$$\begin{aligned}
&\boldsymbol{e}_1^T\left[\boldsymbol{I}_n \mp \boldsymbol{X}_1\left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T\right]\boldsymbol{e}_1 \\
= \;&\boldsymbol{e}_1^T\left[\boldsymbol{I}_n \mp \boldsymbol{X}_1\left\{\boldsymbol{A} \mp \boldsymbol{A}\boldsymbol{X}_1^T (\boldsymbol{I}_n \mp \boldsymbol{H}_1)^{-1}\boldsymbol{X}_1\boldsymbol{A}\right\}\boldsymbol{X}_1^T\right]\boldsymbol{e}_1 \\
= \;&\boldsymbol{e}_1^T\left[\boldsymbol{I}_n \mp \boldsymbol{X}_1\boldsymbol{A}\boldsymbol{X}_1^T + \boldsymbol{X}_1\boldsymbol{A}\boldsymbol{X}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{X}_1\boldsymbol{A}\boldsymbol{X}_1^T\right]\boldsymbol{e}_1 \\
= \;&\boldsymbol{e}_1^T\left[\boldsymbol{I}_n \mp \boldsymbol{H}_1\left\{\boldsymbol{I}_n \mp (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{H}_1\right\}\right]\boldsymbol{e}_1 \\
= \;&\boldsymbol{e}_1^T\left[\boldsymbol{I}_n \mp \boldsymbol{H}_1 (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\right]\boldsymbol{e}_1 \\
= \;&\boldsymbol{e}_1^T (\boldsymbol{I}_n \pm \boldsymbol{H}_1)^{-1}\boldsymbol{e}_1.
\end{aligned}$$

Finally, the first expression for $\boldsymbol{V}_n$ follows from the Matrix Inversion Lemma (Theorem A.4.4) while the second is by definition.

Note that in each pair of equations for the updated quantities, the first equation is appropriate when $n \leq p$ while the second is for $n \geq p$. The sweep operator can be used in either case. For $n \leq p$, we have

$$\begin{aligned} &\text{SWEEP}\,(1,\ldots,p+n)\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}_1^T \\ \mp\boldsymbol{X}_1 & \boldsymbol{I}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{D} & \boldsymbol{E} \end{bmatrix} \\ = &\begin{bmatrix} \boldsymbol{A} \mp \boldsymbol{X}_1^T\left(\boldsymbol{I}_n \pm \boldsymbol{H}_1\right)^{-1}\boldsymbol{X}_1\boldsymbol{A} & -\boldsymbol{A}\boldsymbol{X}_1^T\left(\boldsymbol{I}_n \pm \boldsymbol{H}_1\right)^{-1} \\ \pm\left(\boldsymbol{I}_n \pm \boldsymbol{H}_1\right)^{-1}\boldsymbol{X}_1\boldsymbol{A} & \left(\boldsymbol{I}_n \pm \boldsymbol{H}_1\right)^{-1} \end{bmatrix} \end{aligned}$$

and thus

$$\hat{\boldsymbol{\beta}}_N = \hat{\boldsymbol{\beta}} \mp \boldsymbol{C}\boldsymbol{e}_1, \qquad RSS_N = RSS \pm \boldsymbol{e}_1^T\boldsymbol{E}\boldsymbol{e}_1, \qquad \text{and} \qquad \boldsymbol{V}_N = \boldsymbol{B}.$$

For $n \geq p$

$$\begin{aligned} &\text{SWEEP}\,(1,\ldots p)\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1 & \boldsymbol{X}_1^T\boldsymbol{e}_1 \\ \pm\boldsymbol{e}_1^T\boldsymbol{X}_1 & \boldsymbol{e}_1^T\boldsymbol{e}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{v} \\ \boldsymbol{w}^T & z \end{bmatrix} \\ &\begin{bmatrix} \left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1} & \left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T\boldsymbol{e}_1 \\ \mp\boldsymbol{e}_1^T\boldsymbol{X}_1\left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1} & \boldsymbol{e}_1^T\left\{\boldsymbol{I}_n \mp \boldsymbol{X}_1\left(\boldsymbol{X}^T\boldsymbol{X} \pm \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1\right\}\boldsymbol{e}_1 \end{bmatrix}. \end{aligned}$$

and thus

$$\hat{\boldsymbol{\beta}}_N = \hat{\boldsymbol{\beta}} \pm \boldsymbol{v}, \qquad RSS_N = RXX \pm \boldsymbol{z}, \qquad \text{and} \qquad \boldsymbol{V}_n = \boldsymbol{U}.$$

A.6 Random Vectors and Multivariate Normal Distribution

In elementary statistics, we often avoid treating vectors of random variables since the variables are independent and can be treated separately. In time series analysis, this is usually not the case.

A.6.1 Basic Definitions

We begin our discussion with a series of definitions.

1. A ***d*-dimension random vector $\boldsymbol{X}$** is a vector of d random variables; that is, $\boldsymbol{X} = (X_1, \ldots, X_d)^T$, where $X_1, \ldots, X_d$ are random variables.

2. The **joint cumulative distribution function** (cdf) F_X of the random vector $\boldsymbol{X}$ is given by
$$F_X(\boldsymbol{x}) = F_X(x_1, \ldots, x_d) = \Pr\left(X_1 \leq x_1, \ldots, X_d \leq x_d\right).$$

3. If there exists a function f_X of d variables such that

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_X(\boldsymbol{y})\, d\boldsymbol{y},$$

then $\boldsymbol{X}$ is said to be a **continuous rand vector** and f_X is called the **joint probability density function** (pdf) of $\boldsymbol{X}$.

4. If $\boldsymbol{Y}$ and $\boldsymbol{X}$ are r-dimensional and d-dimensional random vectors, then $\boldsymbol{Z} = (\boldsymbol{Y}^T, \boldsymbol{X}^T)^T$ is an $(r+d)$-dimensional random vector. If f_Z, f_Y, and f_X are the pdfs of $\boldsymbol{Z}, \boldsymbol{Y}$, and $\boldsymbol{X}$, respectively, then the functions

$$f_{Y\,|\,X} = \frac{f_Z}{f_X} \quad \text{and} \quad f_{X\,|\,Y} = \frac{f_Z}{f_Y}$$

are called the **conditional pdfs of $\boldsymbol{Y}$ given $\boldsymbol{X}$** and **$\boldsymbol{X}$ given $\boldsymbol{Y}$**, respectively.

5. The **mean vector $\boldsymbol{\mu}_X$** and **covariance matrix $\boldsymbol{\Sigma}_X$** of $\boldsymbol{X}$ are given by

$$\begin{aligned}
\boldsymbol{\mu}_X &= \mathrm{E}\,[\boldsymbol{X}] = \begin{bmatrix} \mathrm{E}\,[X_1] \\ \vdots \\ \mathrm{E}\,[X_d] \end{bmatrix} = \int \cdots \int \boldsymbol{x} f_X(\boldsymbol{x})\, d\boldsymbol{x}, \\
\boldsymbol{\Sigma}_X &= \mathrm{Var}\,(\boldsymbol{X}) = \mathrm{E}\left[(\boldsymbol{X} - \boldsymbol{\mu}_X)(\boldsymbol{X} - \boldsymbol{\mu}_X)^T\right] \\
&= \int \cdots \int (\boldsymbol{x} - \boldsymbol{\mu}_X)(\boldsymbol{x} - \boldsymbol{\mu}_X)^T f_X(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \begin{bmatrix}
\mathrm{Var}\,(X_1) & \mathrm{Cov}\,(X_1,\, X_2) & \mathrm{Cov}\,(X_1,\, X_3) & \cdots & \mathrm{Cov}\,(X_1,\, X_d) \\
 & \mathrm{Var}\,(X_2) & \mathrm{Cov}\,(X_2,\, X_3) & \cdots & \mathrm{Cov}\,(X_2,\, X_d) \\
 & & \mathrm{Var}\,(X_3) & \cdots & \vdots \\
 & & & \ddots & \vdots \\
 & \text{Symmetric} & & & \mathrm{Var}\,(X_d)
\end{bmatrix}.
\end{aligned}$$

If $\mathrm{E}\,[\boldsymbol{X}] = \boldsymbol{\mu}_X$ and $\mathrm{Var}\,(\boldsymbol{X}) = \boldsymbol{\Sigma}_X$ we will write

$$\boldsymbol{X} \sim (\boldsymbol{\mu}_X,\, \boldsymbol{\Sigma}_X)\,.$$

The covariance matrix $\boldsymbol{\Sigma}_{XY}$ and $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the $d \times r$ matrix whose $(j,\, k)^{\text{th}}$ element is $\mathrm{Cov}\,(X_j,\, Y_k)$. The conditional mean $\boldsymbol{\mu}_{Y\,|\,X}$ and variance $\boldsymbol{\Sigma}_{Y\,|\,X}$ of a random vector $\boldsymbol{Y}$ given $\boldsymbol{X}$ are defined in the same way as $\boldsymbol{\mu}_Y$ and $\boldsymbol{\Sigma}_Y$, using the conditional pdf of $\boldsymbol{Y}$ given $\boldsymbol{X}$.

6. If we partition the $(r+d)$-dimensional random vector $\boldsymbol{Z}$ into the r-dimensional and d-dimensional random vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$ consisting of the first r and last d elements of $\boldsymbol{Z}$, then we can partition $\boldsymbol{\mu}_Z$ and $\boldsymbol{\Sigma}_Z$ similarly as

$$\boldsymbol{\mu}_Z = \begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_Z = \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix},$$

where in fact

$$\begin{aligned} \boldsymbol{\Sigma}_{YY} &= \text{Var}\,(\boldsymbol{Y}), \qquad \boldsymbol{\Sigma}_{YX} = \text{Cov}\,(\boldsymbol{Y}, \boldsymbol{X}) = \boldsymbol{\Sigma}_{XY}^T, \\ \boldsymbol{\Sigma}_{XX} &= \text{Var}\,(\boldsymbol{X}), \qquad \boldsymbol{\Sigma}_{XY} = \text{Cov}\,(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma}_{YX}^T. \end{aligned}$$

7. The r-dimensional random vector $\boldsymbol{Y}$ is called a **linear transformation of the d-dimensional random vector $\boldsymbol{X}$** if $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ where $\boldsymbol{A}$ is an $r \times d$ matrix of constants $\boldsymbol{b}$ is an r-dimensional vector of constants.

8. If $\boldsymbol{X}$ is a d-dimensional random vector and $\boldsymbol{A}$ is a $d \times d$ matrix of constants, then the scalar random variable $Q = \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$ is called a **quadratic form in $\boldsymbol{X}$** and $\boldsymbol{A}$ is called the **matrix of the quadratic form**.

A.6.2 Basic Facts

With these definitions in mind, we next state some theoretical results that are used throughout the book. We begin with the following theorem.

Theorem A.6.1 Mean and Variance of a Linear Function.
If $\boldsymbol{X} \sim (\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ then

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \sim \left(\boldsymbol{A}\boldsymbol{\mu}_X + \boldsymbol{b},\, \boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}^T\right),$$

that is,

$$\begin{aligned} E[\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}] &= \boldsymbol{A}E[\boldsymbol{X}] + \boldsymbol{b} \\ Var(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) &= \boldsymbol{A}\,Var(\boldsymbol{X})\,\boldsymbol{A}^T. \end{aligned}$$

Most of the theory and methods of time series analysis are built upon the properties of the multivariate normal distribution.

Definition A.6.1 Multivariate Normal Distribution.
The d-dimensional random vector $\boldsymbol{X}$ having mean $\boldsymbol{\mu}_X$ and variance $\boldsymbol{\Sigma}_X$ is said to have the **d-dimensional normal distribution** *if its pdf is given by*

$$f_X(\boldsymbol{x}) = (2\pi)^{-d/2}\,|\boldsymbol{\Sigma}_X|^{-1/2} \exp\left\{-\frac{1}{2}\,(\boldsymbol{x} - \boldsymbol{\mu}_X)^T\,\boldsymbol{\Sigma}_X^{-1}\,(\boldsymbol{x} - \boldsymbol{\mu}_X)\right\},$$

for $\boldsymbol{x} \in \Re^d$, that is, in d-dimensional Euclidean space. Such a random vector is denoted by $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$.

The important properties of the multivariate normal distribution are summaries in Theorem A.6.2 (see Rao (1973), Chapter 8, for example).

Theorem A.6.2 PROPERTIES OF THE MULTIVARIATE NORMAL.
If

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix} \sim N_{r+d}\left(\begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix} \right),$$

then

(a) The characteristic function of $\boldsymbol{Z}$ is given by

$$\phi_Z(\boldsymbol{t}) = E\left[e^{i\boldsymbol{t}^T\boldsymbol{Z}}\right] = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu}_Z - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_Z\boldsymbol{t}\right).$$

(b) If $\boldsymbol{A}$ is an $s \times (r+d)$ matrix of constants and $\boldsymbol{b}$ is an s-dimensional vector of constants, then

$$\boldsymbol{A}\boldsymbol{Z} + \boldsymbol{b} \sim N_s\left(\boldsymbol{A}\boldsymbol{\mu}_Z + \boldsymbol{b},\, \boldsymbol{A}\boldsymbol{\Sigma}_Z\boldsymbol{A}^T\right);$$

that is, linear transformations of normal random vectors are also normally distributed.

(c) (i) The conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim N_r\left(\boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X),\, \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\right);$$

that is,the pdf of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is that of an r-dimensional normal random vector with mean and variance given by

$$\begin{aligned} \boldsymbol{\mu}_{Y\mid X} &= E[\boldsymbol{Y} \mid \boldsymbol{X}] = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X) \\ \boldsymbol{\Sigma}_{Y\mid X} &= Var(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}; \end{aligned}$$

that is the condition mean of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ is a linear function of $\boldsymbol{x}$, while the conditional variance is the same, no matter what the value of $\boldsymbol{x}$ is.

(ii) Conversely, if

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim N_r\left(\boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X),\, \boldsymbol{\Sigma}_{YY} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\right),$$

and $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}_X,\, \boldsymbol{\Sigma}_{XX})$, then

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix} \sim N_{r+d}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix} \right).$$

(d) If $\boldsymbol{A}$ is a $d \times d$ nonsingular matrix of constants, then the distribution of $\boldsymbol{Y} \mid \boldsymbol{X}$ and $\boldsymbol{Y} \mid \boldsymbol{A}\boldsymbol{X}$ are the same.

(e) $Cov(\boldsymbol{A}\boldsymbol{X},\, \boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\Sigma}_{XY}\boldsymbol{B}^T$.

(f) The quadratic form

$$Q = (\boldsymbol{X} - \boldsymbol{\mu}_X)^T\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X)$$

has the chi-square distribution with d degrees of freedom, and we write $Q \sim \chi_d^2$.

(g) The quadratic form

$$Q_A = (\boldsymbol{X} - \boldsymbol{\mu}_X)^T \boldsymbol{A} (\boldsymbol{X} - \boldsymbol{\mu}_X)$$

has the same distribution as the random variable $\sum_{j=1}^{d} \lambda_j Z_j$, where $Z_1, \ldots, Z_s$ are independent χ^2_1 random variables, and the λ_i are the non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{\Sigma}_{XX}$. This reduces to the result of part (f) if $\boldsymbol{A} = \boldsymbol{\Sigma}_{XX}^{-1}$ since the eigenvalues of the d-dimensional identity matrix are just one repeated d times, and the sum of d independent χ^2_1 random variables is χ^2_d.

(h) The mean and variance of a quadratic form in normal variables are given by

$$\begin{aligned} E\left[\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}\right] &= tr(\boldsymbol{A}\boldsymbol{\Sigma}_{XX}) + \boldsymbol{\mu}_X^T \boldsymbol{A} \boldsymbol{\mu}_X \\ Var\left(\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}\right) &= 2tr\left([\boldsymbol{A}\boldsymbol{\Sigma}_{XX}]^2\right) + 4\boldsymbol{\mu}_X^T \boldsymbol{A}\boldsymbol{\Sigma}_{XX}\boldsymbol{A}^T \boldsymbol{\mu}_X. \end{aligned}$$

(i) The two quadratic forms $Q_1 = \boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}$ and $Q_2 = \boldsymbol{X}^T \boldsymbol{B} \boldsymbol{X}$ are independent if and only if $\boldsymbol{A}\boldsymbol{\Sigma}_{XX}\boldsymbol{B}^T = \boldsymbol{0}$.

(j) If Q_1 and Q_2 are two independent χ^2 random variables having n_1 and n_2 degrees of freedom, then the random variable

$$F = \frac{Q_1/n_1}{Q_2/n_2}$$

has the F distribution with n_1 numerator degrees of freedom and n_2 denominator degrees of freedom, and we write $F \sim F_{n_1,n_2}$.

(k) If $Z \sim N_1(\mu, \sigma^2)$, and $Q \sim \chi^2_v$ are independent, then the random variable

$$t = \frac{(Z - \mu)/\sigma}{\sqrt{Q/v}}$$

has the Student's t distribution with v degrees of freedom and we write $t \sim t_v$.

Computing the pdf, cdf, and Quantile Functions

The *R* software has functions for computing the pdf, cdf, and quantile functions for the $\mathrm{N}(\mu\,\sigma^2)$, t_v, χ^2_v, and F_{v_1,v_2} distributions. They are

For the normal distribution with $\mu =$ `mean` and standard deviation $\sigma =$ `sd`, the pdf at `x` is computed using `dnorm(x, mean, sd)`, the cdf at `x` is computed using `pnorm(x, mean, sd)`, and the p^{th} quantile is computed using `qnorm(p, mean, sd)`. For all three functions, the default values for the mean and standard deviation are `mean = 0` and `sd = 1`.

- For the Student's t_v distribution with degrees of freedom $v =$ `df`, the pdf at `x` is computed using `dt(x, df)`, the cdf at `x` is computed using `pt(x, df)`, and the p^{th} quantile is computed using `qt(p, df)`.

- For the χ^2_v with degrees of freedom $v =$ `df`, the pdf at `x` is computed using `dchisq(x, df)`, the cdf at `x` is computed using `pchisq(x, df)`, and the p^{th} quantile is computed using `qchisq(p, df)`.
- For the F_{v_1,v_2} with numerator degrees of freedom $v_1 =$ `df1`, and denominator degrees of freedom $v_2 =$ `df2` the pdf at `x` is computed using `df(x, df1, df2)`, the cdf at `x` is computed using `pf(x, df1, df2)`, and the p^{th} quantile is computed using `qf(p, df1, df2)`.

These functions may be computed for many other distributions in R using the same type of syntax. For a comprehensive list and documentations, issue the command `help("Distributions")` at the R prompt.

Simultaneous Confidence Intervals

A result that is useful in finding simultaneous confidence bands for functions is given in the next theorem, due to Scheffé (1959), p. 406.

Theorem A.6.3 Simultaneous Confidence Intervals.
Let $\boldsymbol{X}$ be an arbitrary p-dimensional random vector and let $\boldsymbol{V}$ be a positive definite $p \times p$ matrix of constants. Then the p-dimensional vector $\boldsymbol{a}$ satisfies

$$(\boldsymbol{X} - \boldsymbol{a})^T \boldsymbol{V} (\boldsymbol{X} - \boldsymbol{a}) \leq 1$$

if and only if

$$\left|\boldsymbol{h}^T (\boldsymbol{X} - \boldsymbol{a})\right| \leq \left(\boldsymbol{h}^T \boldsymbol{V}^{-1} \boldsymbol{h}\right)^{-1/2} = D$$

for all p-dimensional vectors $\boldsymbol{h}$; that is,

$$\boldsymbol{h}^T \boldsymbol{X} - D \leq \boldsymbol{h}^T \boldsymbol{a} \leq \boldsymbol{h}^T \boldsymbol{X} + D.$$

A.6.3 Prediction of a Random Vector

Consider random vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$ such that $\boldsymbol{Y}$ has d elements, $\boldsymbol{X}$ has r elements, and

$$\mathrm{E}\left[\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix}\right] = \begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{bmatrix} \quad \mathrm{Var}\left(\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{bmatrix}.$$

We consider to criteria for predicting $\boldsymbol{Y}$ from $\boldsymbol{X}$: (1) best unbiased prediction and (2) best linear unbiased prediction.

Definition A.6.2 Best Unbiased Predictor & Best Linear Unbiased Predictor.
The **best unbiased predictor of $\boldsymbol{Y}$ from $\boldsymbol{X}$** *is that function $\tilde{\boldsymbol{Y}}$ of $\boldsymbol{X}$ that has the same expectation as $\boldsymbol{Y}$ and is closest to $\boldsymbol{Y}$ in expected mean square; that is,*

$$E\left[(\boldsymbol{Y} - \boldsymbol{Y}^*)(\boldsymbol{Y} - \boldsymbol{Y}^*)^T\right] - E\left[\left(\boldsymbol{Y} - \tilde{\boldsymbol{Y}}\right)\left(\boldsymbol{Y} - \tilde{\boldsymbol{Y}}\right)^T\right]$$

is positive semidefinite for any other unbiased function $\boldsymbol{Y}^*$ *of* $\boldsymbol{X}$. *The* **best linear unbiased predictor** $\hat{\boldsymbol{Y}}$ *of* $\boldsymbol{Y}$ *given* $\boldsymbol{X}$ *is that linear function of* $\boldsymbol{X}$ *having the same expectation of* $\boldsymbol{Y}$ *and is closest to* $\boldsymbol{Y}$ *in mean squared error; that is,* $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{A}}\boldsymbol{X}$ *where*

$$E\left[(\boldsymbol{Y} - \boldsymbol{A}^*\boldsymbol{X})(\boldsymbol{Y} - \boldsymbol{A}^*\boldsymbol{X})^T\right] - E\left[\left(\boldsymbol{Y} - \hat{\boldsymbol{A}}\boldsymbol{X}\right)\left(\boldsymbol{Y} - \hat{\boldsymbol{A}}\boldsymbol{X}\right)^T\right]$$

is positive semidefinite for any other linear function $\boldsymbol{A}^*\boldsymbol{X}$ *that is also unbiased for* $\boldsymbol{Y}$.

It will turn out that $\tilde{\boldsymbol{Y}}$ is the conditional expectation of $\boldsymbol{Y}$ given $\boldsymbol{X}$. Conditional expectation is a crucial concept in time series analysis and can be defined at many levels of generality. A more general and useful definition of conditional expectation that is given in Definition 5 of Appendix A.6.1 is given by the following (see Parzen (1962b), for example).

Definition A.6.3 Conditional Expectation of a Random Vector.
Let $\boldsymbol{Y}$ *be a random vector having finite mean. Then the* **conditional expectation** $\mathbf{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ **of** $\boldsymbol{Y}$ **given** $\boldsymbol{X}$ *is that function* ϕ *of* $\boldsymbol{X}$ *such that*

$$E[\{\boldsymbol{Y} - \phi(\boldsymbol{X})\}\boldsymbol{g}(\boldsymbol{X})] = \mathbf{0}$$

for any bounded function $\boldsymbol{g}$ *of* $\boldsymbol{X}$.

This definition is called descriptive (as opposed to the constructive definition of Appendix A.6.1) as it says that the "residual" $\boldsymbol{Y} - \mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ is uncorrelated with any other function of $\boldsymbol{X}$. In this sense, there is no more information about $\boldsymbol{Y}$ in $\boldsymbol{X}$ after having removed $\mathrm{E}[\boldsymbol{Y} \mid bX]$. We note that the conditional variance $\mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{X})$ of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is just the conditional expectation of the random variable $\left[(\boldsymbol{Y} - \mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{X}])(\boldsymbol{Y} - \mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{X}])^T\right]$ given $\boldsymbol{X}$.

Letting $\boldsymbol{g}(\boldsymbol{X}) = \mathbf{1}$ in this definition yields $\mathrm{E}[\phi(\boldsymbol{X})] = \mathrm{E}[\boldsymbol{Y}]$, and thus

$$\mathrm{Cov}(\boldsymbol{Y} - \phi(\boldsymbol{X}), \boldsymbol{g}(\boldsymbol{X})) = \mathbf{0}.$$

With these definitions in mind, we have the following theorem. For simplicity, we assume that the random vectors involved have zero mean.

Theorem A.6.4 General Prediction Theory.
Let $\boldsymbol{Y}$ *and* $\boldsymbol{X}$ *be jointly distributed random vectors having zero mean. Then*

(a) The best unbiased predictor $\tilde{\boldsymbol{Y}}$ *of* $\boldsymbol{Y}$ *given* $\boldsymbol{X}$ *and its error variance* $\boldsymbol{\Sigma}_{\tilde{Y}} = Var\left(\boldsymbol{Y} - \tilde{\boldsymbol{Y}}\right)$ *are given by*

$$\tilde{\boldsymbol{Y}} = E[\boldsymbol{Y} \mid \boldsymbol{X}] \quad \text{and} \quad \boldsymbol{\Sigma}_{\tilde{Y}} = Var(\boldsymbol{Y} \mid \boldsymbol{X}).$$

(b) If $\boldsymbol{\Sigma}_{XX}$ *is nonsingular, then the best linear unbiased predictor* $\hat{\boldsymbol{Y}}$ *of* $\boldsymbol{Y}$ *given* $\boldsymbol{X}$ *and its error variance* $\boldsymbol{\Sigma}_{\hat{Y}} = Var\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)$ *are given by*

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{A}}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{\Sigma}_{\hat{Y}} = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY},$$

where $\hat{\boldsymbol{A}}$ *satisfies the prediction normal equations*

$$\boldsymbol{\Sigma}_{XX}\hat{\boldsymbol{A}}^T = \boldsymbol{\Sigma}_{XY}.$$

(c) If $\boldsymbol{Y}$ and $\boldsymbol{X}$ are jointly multivariate normal, then $\tilde{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}$ and $\boldsymbol{\Sigma}_{\tilde{Y}} = \boldsymbol{\Sigma}_{\hat{Y}}$.

Proof: To prove part (a), let $\boldsymbol{Y}^*$ be an arbitrary function of $\boldsymbol{X}$ having mean $\mathbf{0}$, and let

$$\boldsymbol{\delta} = \boldsymbol{Y} - \mathrm{E}\left[\boldsymbol{Y} \mid \boldsymbol{X}\right], \quad \boldsymbol{\gamma} = \mathrm{E}\left[\boldsymbol{Y} \mid \boldsymbol{X}\right] - \boldsymbol{Y}^*, \quad \text{and} \quad \boldsymbol{\alpha} = \boldsymbol{Y} - \boldsymbol{Y}^* = \boldsymbol{\delta} + \boldsymbol{\gamma}.$$

We need to show that $\mathrm{E}\left[\boldsymbol{\alpha}\boldsymbol{\alpha}^T\right] - \mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}^T\right]$ is positive semidefinite. We have

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{\alpha}\boldsymbol{\alpha}^T\right] - \mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}^T\right] &= \mathrm{E}\left[(\boldsymbol{\delta}+\boldsymbol{\gamma})(\boldsymbol{\delta}+\boldsymbol{\gamma})^T\right] - \mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}^T\right] \\ &= \mathrm{E}\left[\boldsymbol{\gamma}\boldsymbol{\delta}^T\right] + \mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\gamma}^T\right] + \mathrm{E}\left[\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right], \end{aligned}$$

the third term of which is positive semidefinite since for any vector $\boldsymbol{h}$ not identically zero, we have

$$\boldsymbol{h}^T \mathrm{E}\left[\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right]\boldsymbol{h} = \mathrm{E}\left[\boldsymbol{h}^T\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{h}\right] = \mathrm{E}\left[\left(\boldsymbol{h}^T\boldsymbol{\gamma}\right)^2\right] \geq 0.$$

Now $\mathrm{E}\left[\boldsymbol{\gamma}\right] = \mathrm{E}\left[\boldsymbol{\delta}\right] = \mathbf{0}$ and thus $\mathrm{E}\left[\boldsymbol{\gamma}\boldsymbol{\delta}^T\right] = \mathrm{Cov}\left(\boldsymbol{\gamma}, \boldsymbol{\delta}\right)$ and $\mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\gamma}^T\right] = \mathrm{Cov}\left(\boldsymbol{\delta}, \boldsymbol{\gamma}\right)$. But both of these covariances are zero since $\boldsymbol{\gamma} = \mathrm{E}\left[\boldsymbol{Y} \mid \boldsymbol{X}\right] - \boldsymbol{Y}^*$ is a function of $\boldsymbol{X}$ since $\mathrm{E}\left[\boldsymbol{Y} \mid \boldsymbol{X}\right]$ and $\boldsymbol{Y}^*$ are both functions of $\boldsymbol{X}$.

To prove part (b), let $\boldsymbol{A}$ be an arbitrary $r \times d$ matrix. Then

$$\begin{aligned} S(\boldsymbol{A}) &= \mathrm{E}\left[(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X})^T\right] \\ &= \mathrm{E}\left[\boldsymbol{Y}\boldsymbol{Y}^T\right] - \boldsymbol{A}\mathrm{E}\left[\boldsymbol{X}\boldsymbol{Y}^T\right] - \mathrm{E}\left[\boldsymbol{Y}\boldsymbol{X}^T\right]\boldsymbol{A}^T + \boldsymbol{A}\mathrm{E}\left[\boldsymbol{X}\boldsymbol{X}^T\right]\boldsymbol{A}^T \\ &= \boldsymbol{\Sigma}_{YY} - \boldsymbol{A}\boldsymbol{\Sigma}_{XY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{\Sigma}_{XX}\boldsymbol{A}^T \\ &= \left(\boldsymbol{A} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\right)\boldsymbol{\Sigma}_{XX}\left(\boldsymbol{A} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\right)^T \\ &\qquad \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}. \end{aligned}$$

Thus $\boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}} = S(\hat{\boldsymbol{A}}) = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$ and

$$S(\boldsymbol{A}^*) - S(\hat{\boldsymbol{A}}) = \left(\boldsymbol{A}^* - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\right)\boldsymbol{\Sigma}_{XX}\left(\boldsymbol{A}^* - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\right)^T$$

which is clearly positive definite for any matrix $\boldsymbol{A}^*$.

Finally, part (c) follows from part c of Theorem A.6.2.

The Kalman Filter Algorithm

If we have a sequence of random vectors that are multivariate normal and satisfy a recursive relationship of a particular form, then predictors can also be found recursively. This type of recursion was used extensively by Kalman (1960) and thus the resulting algorithm is usually referred to as the Kalman Filter Algorithm.

Theorem A.6.5 THE KALMAN FILTER.
Let $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ and $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots$ be sequences of r- and d-dimensional random vectors satisfying the two equations (called the **state** *and* **observation equations** *respectively) for $t \geq 1$*

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{A}_t \boldsymbol{X}_{t-1} + \boldsymbol{w}_t \\ \boldsymbol{Z}_t &= \boldsymbol{H}_t \boldsymbol{X}_t + \boldsymbol{v}_t, \end{aligned}$$

where the $\boldsymbol{A}_t$ and $\boldsymbol{H}_t$ are full rank $r \times r$ and $d \times r$ matrices of constants,

$$\begin{aligned} \boldsymbol{w}_t &\sim N_r(\boldsymbol{0}_r, \boldsymbol{W}_t) \\ \boldsymbol{v}_t &\sim N_d(\boldsymbol{0}_d, \boldsymbol{V}_t) \end{aligned}$$

all of the $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are independent, and

$$\boldsymbol{X}_0 \sim N_r(\boldsymbol{\mu}_0, \boldsymbol{\Gamma}_0).$$

Let $\hat{\boldsymbol{X}}_t$ and $\hat{\boldsymbol{\Sigma}}_t$ denote the mean and variance of the conditional distribution of $\boldsymbol{X}_t$ given $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t$. Then

(a) For $t \geq 1$

$$\begin{aligned} \hat{\boldsymbol{X}}_t &= \boldsymbol{A}_t \hat{\boldsymbol{X}}_{t-1} + \boldsymbol{R}_t \boldsymbol{H}_t^T \left(\boldsymbol{H}_t \boldsymbol{R}_t \boldsymbol{H}_t^T + \boldsymbol{V}_t \right)^{-1} \boldsymbol{e}_t \\ \hat{\boldsymbol{\Sigma}}_t &= \boldsymbol{R}_t - \boldsymbol{R}_t \boldsymbol{H}_t^T \left(\boldsymbol{H}_t \boldsymbol{R}_t \boldsymbol{H}_t^T + \boldsymbol{V}_t \right)^{-1} \boldsymbol{H}_t \boldsymbol{R}_t, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{R}_t &= \boldsymbol{A}_t \hat{\boldsymbol{\Sigma}}_{t-1} \boldsymbol{A}_t^T + \boldsymbol{W}_t \\ \boldsymbol{e}_t &= \boldsymbol{Z}_t - \boldsymbol{H}_t \boldsymbol{A}_t \hat{\boldsymbol{X}}_{t-1}, \end{aligned}$$

and

$$\hat{\boldsymbol{\Sigma}}_0 = \boldsymbol{\Sigma}_0 \quad \textit{and} \quad \hat{\boldsymbol{X}}_0 = \boldsymbol{\mu}_0.$$

(b) For $t > 1$ and $h \geq 0$, let $\tilde{\boldsymbol{X}}_{t+h-1\,|\,t-1}$ and $\tilde{\boldsymbol{\Sigma}}_{t+h-1\,|\,t-1}$ denote the mean and variance of $\boldsymbol{X}_{t+h-1}$ given $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{t-1}$. Then

$$\begin{aligned} \tilde{\boldsymbol{X}}_{t+h-1\,|\,t-1} &= \begin{cases} \hat{\boldsymbol{X}}_{t-1}, & h = 0 \\ \boldsymbol{A}_{t+h-1} \tilde{\boldsymbol{X}}_{t+h-2\,|\,t-1}, & h \geq 1 \end{cases} \\ \tilde{\boldsymbol{\Sigma}}_{t+h-1\,|\,t-1} &= \begin{cases} \hat{\boldsymbol{\Sigma}}_{t-1}, & h = 0 \\ \boldsymbol{A}_{t+h-1} \tilde{\boldsymbol{\Sigma}}_{t+h-2\,|\,t-1} \boldsymbol{A}_{t+h-1}^T + \boldsymbol{W}_t, & h \geq 1. \end{cases} \end{aligned}$$

Proof: In our proof, we follow Meinhold & Singpurwalla (1983). Let $\boldsymbol{Z}^t$ represent the set of variables $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t\}$. We know that the distribution of $\boldsymbol{X}_t \,|\, \boldsymbol{Z}^t$, which is what we want, is normal since $\boldsymbol{X}_0$ and all of the $\boldsymbol{w}_t$s and $\boldsymbol{v}_t$s are normal, and thus all of the $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$

are jointly normal. We want to show that the mean and variance of $\boldsymbol{X}_t \,|\, \boldsymbol{Z}^t$ satisfy the relations in part (a). Note that $\hat{\boldsymbol{X}}_{t-1} = \mathrm{E}\left[\boldsymbol{X}_{t-1} \,|\, \boldsymbol{Z}^{t-1}\right]$ is a linear function of $\boldsymbol{Z}^{t-1}$ and thus $\boldsymbol{e}_t = \boldsymbol{Z}_t - \boldsymbol{H}_t\boldsymbol{Z}_t\hat{\boldsymbol{X}}_{t-1}$ and $\boldsymbol{Z}^{t-1}$ are a linear transformation of $\boldsymbol{Z}^t$. Thus, by part (d) of Theorem A.6.2, we know that the conditional distribution of $\boldsymbol{X}_t \,|\, \boldsymbol{Z}^t$ is the same as that of $\boldsymbol{X}_t \,|\, \boldsymbol{e}_t, \boldsymbol{Z}^{t-1}$. The distribution of $\boldsymbol{X}_t \,|\, \boldsymbol{e}_t, \boldsymbol{Z}^{t-1}$ can be found from that of $\begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{e}_t \end{pmatrix} \,|\, \boldsymbol{Z}^{t-1}$ by conditioning on $\boldsymbol{e}_t$ (see part (ci) of Theorem A.6.2), and the distribution of $\begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{e}_t \end{pmatrix} \,|\, \boldsymbol{Z}^{t-1}$ can be found from $\boldsymbol{X}_t \,|\, \boldsymbol{Z}^{t-1}$ and $\boldsymbol{e}_t \,|\, \boldsymbol{X}_t, \boldsymbol{Z}^{t-1}$ by part (cii) of Theorem A.6.2. When $t = 1$ we want the distribution of $\boldsymbol{X}_1 \,|\, \boldsymbol{Z}_1$, so we argue in the same way except there is no $\boldsymbol{Z}^{t-1}$.

Since $\boldsymbol{X}_t = \boldsymbol{A}_t\boldsymbol{X}_{t-1} + \boldsymbol{w}_t$, we have

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{X}_t \,|\, \boldsymbol{Z}^{t-1}\right] &= \boldsymbol{A}_t\mathrm{E}\left[\boldsymbol{X}_{t-1} \,|\, \boldsymbol{Z}^{t-1}\right] + \mathrm{E}\left[\boldsymbol{w}_t \,|\, \boldsymbol{Z}^{t-1}\right] = \boldsymbol{A}_t\hat{\boldsymbol{X}}_{t-1} \\ \mathrm{Var}\left(\boldsymbol{X}_t \,|\, \boldsymbol{Z}^{t-1}\right) &= \boldsymbol{A}_t\mathrm{Var}\left(\boldsymbol{X}_{t-1} \,|\, \boldsymbol{Z}^{t-1}\right)\boldsymbol{A}_t^T + \mathrm{Var}\left(\boldsymbol{w}_t \,|\, \boldsymbol{Z}^{t-1}\right) = \boldsymbol{A}_t\hat{\boldsymbol{\Sigma}}_{t-1}\boldsymbol{A}_t^T + \boldsymbol{W}_t, \end{aligned}$$

that is

$$\boldsymbol{X}_t \,|\, \boldsymbol{Z}^{t-1} \sim N_r(\boldsymbol{A}_t\hat{\boldsymbol{X}}_{t-1},\, \boldsymbol{A}_t\hat{\boldsymbol{\Sigma}}_{t-1}\boldsymbol{A}_t^T + \boldsymbol{W}_t).$$

Since $\boldsymbol{e}_t = \boldsymbol{Z}_t - \boldsymbol{H}_t\boldsymbol{A}_t\hat{\boldsymbol{X}}_{t-1}$ and $\boldsymbol{Z}_t = \boldsymbol{H}_t\boldsymbol{X}_t + \boldsymbol{v}_t$, we have $\boldsymbol{e}_t = \boldsymbol{H}_t\left[bX_t - \boldsymbol{A}_t\hat{\boldsymbol{X}}_{t-1}\right] + \boldsymbol{v}_t$, and since $\hat{\boldsymbol{X}}_{t-1}$ is a function of $\boldsymbol{Z}^{t-1}$, we have

$$\boldsymbol{e}_t \,|\, \boldsymbol{X}_t, \boldsymbol{Z}^{t-1} \sim N_d\left(\boldsymbol{H}_t\left\{\boldsymbol{X}_t - \boldsymbol{A}_t\hat{\boldsymbol{X}}_{t-1}\right\}, \boldsymbol{V}_t\right).$$

Let $\boldsymbol{R}_t = \boldsymbol{A}_t\hat{\boldsymbol{\Sigma}}_{t-1}\boldsymbol{A}_t^T + \boldsymbol{W}_t$, and to use part (cii) of Theorem A.6.2, make the identifications

$$\boldsymbol{X} \leftrightarrow \boldsymbol{e}_t, \qquad \boldsymbol{Y} \leftrightarrow \boldsymbol{X}_t, \qquad \boldsymbol{\mu}_Y \leftrightarrow \boldsymbol{A}_t\hat{\boldsymbol{X}}_{t-1}, \qquad \boldsymbol{\mu}_X \leftrightarrow \boldsymbol{0},$$

$$\boldsymbol{\Sigma}_{XX} \leftrightarrow \boldsymbol{H}_t\boldsymbol{R}_T\boldsymbol{H}_t^T + \boldsymbol{V}_t, \qquad \boldsymbol{\Sigma}_{XX} \leftrightarrow \boldsymbol{H}_t\boldsymbol{R}_t, \qquad \boldsymbol{\Sigma}_{YY} \leftrightarrow \boldsymbol{R}_t.$$

To verify that this identification is valid, we suppress reference to $\boldsymbol{Z}^{t-1}$ and note that

$$\begin{aligned} \boldsymbol{e}_t \,|\, \boldsymbol{X}_t &\leftrightarrow \boldsymbol{X} \,|\, \boldsymbol{Y} \\ &\sim N_d\left(\boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\left\{\boldsymbol{Y} - \boldsymbol{\mu}_T\right\}, \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\right) \\ &= N_d\left(\boldsymbol{H}_t\boldsymbol{R}_t\boldsymbol{R}_t^{-1}\left\{\boldsymbol{X}_t - \boldsymbol{A}_t\hat{\boldsymbol{X}}_{t-1}\right\}, \boldsymbol{H}_t\boldsymbol{R}_t\boldsymbol{H}_t^T + \boldsymbol{V}_t - \boldsymbol{H}_t\boldsymbol{R}_t\boldsymbol{R}_t^{-1}\boldsymbol{R}_t\boldsymbol{H}_t^T\right) \\ &= N_d\left(\boldsymbol{H}_t\left\{\boldsymbol{X}_t - \boldsymbol{A}\hat{\boldsymbol{X}}_{t-1}\right\}, \boldsymbol{V}_t\right) \end{aligned}$$

as required. Thus we have by part (cii) of Theorem A.6.2,

$$\begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{e}_t \end{pmatrix} \,\Big|\, \boldsymbol{Z}^{t-1} \sim N_{r+d}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{A}_t\hat{\boldsymbol{X}}_{t-1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{H}_t\boldsymbol{R}_t\boldsymbol{H}_t^T + \boldsymbol{V}_t & \boldsymbol{H}_t\boldsymbol{R}_t \\ \boldsymbol{R}_t^T\boldsymbol{H}_t^T & \boldsymbol{R}_t \end{bmatrix},\right)$$

which in turn gives the desired recursion for $\hat{\boldsymbol{X}}_t$ and $\hat{\boldsymbol{\Sigma}}_t$ by part (ci) of Theorem A.6.2.

Part (b) for $h = 0$ follows by definition, while $h > 0$ the results come from the fact that $\boldsymbol{X}_{t+h-1} = \boldsymbol{A}_{t+h-2}\boldsymbol{X}_{t+h-2} + \boldsymbol{w}_{t+h-1}$.

Alternate Expression of the Kalman Filter Algorithm

There are many algebraically equivalent ways to express the Kalman Filter Algorithm (KFA). In Theorem A.6.5 we wrote it in a way making it clear how the recursion goes from one step the next, and how it starts. To obtain a more intuitively appealing expressions, define

$$\tilde{\boldsymbol{X}}_t = \mathrm{E}\left[\boldsymbol{X}_t \,|\, \boldsymbol{Z}^{t-1}\right] \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}_t = \mathrm{Var}\left(\boldsymbol{X}_t \,|\, \boldsymbol{Z}^{t-1}\right),$$

that is, the mean and variance of $\boldsymbol{X}_t$ having only observed $\boldsymbol{Z}$ up to time $t-1$. Then clearly

$$\tilde{\boldsymbol{X}}_t = \boldsymbol{A}_t \hat{\boldsymbol{X}}_{t-1} \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}_t = \boldsymbol{R}_t;$$

that is, our best "estimate" of $\boldsymbol{X}_t$ without having observed $\boldsymbol{Z}_t$ is $\boldsymbol{A}_T \hat{\boldsymbol{X}}_{t-1}$. If we define the Kalman gain matrix $\boldsymbol{K}_t$ by

$$\boldsymbol{K}_t = \tilde{\boldsymbol{\Sigma}}_t \boldsymbol{H}_t^T \left(\boldsymbol{H}_t \tilde{\boldsymbol{\Sigma}}_t \boldsymbol{H}_t^T + \boldsymbol{V}_t\right)^{-1},$$

we can write

$$\hat{\boldsymbol{X}}_t = \tilde{\boldsymbol{X}}_t + \boldsymbol{K}_t \left(\boldsymbol{Z}_t - \boldsymbol{H}_t \tilde{\boldsymbol{X}}_t\right);$$

that is, our best guess of $\boldsymbol{X}_t$ given $\boldsymbol{Z}^t$ is a linear combination of $\tilde{\boldsymbol{X}}_t$ (our best guess of $\boldsymbol{X}_t$ given $\boldsymbol{Z}^{t-1}$) and a variable expressing how well $\boldsymbol{H}_t \tilde{\boldsymbol{X}}_t$ predicts the new observation $\boldsymbol{Z}_t$. Note hoe similar this is to recursive regression (see Theorem A.5.1). We can summarize the alternative expression for the KFA as $\hat{\boldsymbol{X}}_0 = \mathrm{E}\left[\boldsymbol{X}_0\right]$, $\hat{\boldsymbol{\Sigma}}_0 = \mathrm{Var}\left(\boldsymbol{X}_0\right)$ and

$$\hat{\boldsymbol{X}}_t = \tilde{\boldsymbol{X}}_t + \boldsymbol{K}_t \boldsymbol{e}_t \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_t = \left[\boldsymbol{I}_t - \boldsymbol{K}_t \boldsymbol{H}_t\right] \tilde{\boldsymbol{\Sigma}}_t,$$

where

$$\begin{aligned}
\tilde{\boldsymbol{X}}_t &= \boldsymbol{A}_t \hat{\boldsymbol{X}}_{t-1} \\
\boldsymbol{e}_t &= \boldsymbol{Z}_t - \boldsymbol{H}_t \tilde{\boldsymbol{X}}_t \\
\tilde{\boldsymbol{\Sigma}}_t &= \boldsymbol{A}_t \hat{\boldsymbol{\Sigma}}_{t-1} \boldsymbol{A}_t^T + \boldsymbol{W}_t \\
\boldsymbol{K}_t &= \tilde{\boldsymbol{\Sigma}}_t \boldsymbol{H}_t^T \left(\boldsymbol{H}_t \tilde{\boldsymbol{\Sigma}}_t \boldsymbol{H}_t^T + \boldsymbol{V}_t\right)^{-1}.
\end{aligned}$$

A.6.4 Convergence of Random Variables

Many of the theoretical results for statistics calculated in time series analysis are stated as limiting results as the length of the time series gets large. In this section we discuss the various modes of convergence of random variables.

Definition A.6.4 TYPES OF CONVERGENCE. *Let $\{X_n,\ n = 1, 2, \ldots\}$ be a sequence of random variables such that the cdf of X_n is F_n. Let X be another random variable, and suppose that the cdf of X is F. Then*

(a) X_n *converges to the constant* c **in probability** *if for every* $\epsilon > 0$,

$$\lim_{n\to\infty} Pr(|X_n - c| > \epsilon) = 0.$$

We denote this by $X_n \xrightarrow{\mathcal{P}} c$, *and say that* $X_n \xrightarrow{\mathcal{P}} X$ *if* $X_n - X \xrightarrow{\mathcal{P}} 0$.

(b) X_n *converges to the constant* c **with probability one** *(or* **almost surely***) if for every* $\epsilon > 0$,

$$\lim_{N\to\infty} Pr\left(\sup_{n\geq N} |X_n - c| > \epsilon\right) = 0.$$

We denote this by $X_n \xrightarrow{a.s.} c$, *and say that* $X_n \xrightarrow{a.s.} X$ *if* $X_n - X \xrightarrow{a.s.} 0$.

(c) X_n *converges to* X **in mean square** *if*

$$\lim_{n\to\infty} E\left[|X_n - X|^2\right] = 0.$$

We denote this by $X_n \xrightarrow{m.s.} X$.

(d) X_n *converges to* X **in distribution** *if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every x *where* F *is continuous. We denote this by* $X_n \xrightarrow{\mathcal{L}} X$.

(e) If $\{Z_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ *is a doubly infinite sequence of random variables and* $\beta_j,\ j = 0, \pm 1, \pm 2, \ldots\}$ *is a sequence of constants, we say that*

$$Y_n = \sum_{j=-\infty}^{\infty} \beta_j Z_{n-j}$$

is a **limit in mean square** *if*

$$Y_{n,M} = \sum_{j=-M}^{M} \beta_j Z_{n-j} \xrightarrow{m.s.} Y_n.$$

We note that a statistic $\hat{\boldsymbol{\theta}}_n$, calculated from a realization of length n, is said to be a weakly (strongly) consistent estimator of a parameter θ if $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathcal{P}} \theta$ ($\hat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \theta$). Also, a random variable X_n is said to be asymptotically normally distributed with asymptotic mean μ_n and asymptotic variance σ_n^2 if

$$\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{\mathcal{L}} X \sim N(0, 1).$$

We denote such a random variable by $X_n \sim AN(\mu_n, \sigma_n^2)$. For example, the usual Central Limit Theorem of statistics says that $\bar{X}_n \sim AN(\mu, n^{-1}\sigma^2)$.

Convergence of Random Vectors

For a sequence of random vectors $\{\boldsymbol{X}_n\}$, we can define concepts of convergence analogous to those for a sequence of single random variables. Converges in probability, with probability one, and in mean square are defined pointwise; that is, we say that $\boldsymbol{X}_n \xrightarrow{\mathcal{P}} \boldsymbol{X}$, $\boldsymbol{X}_n \xrightarrow{a.s.} \boldsymbol{X}$, or $\boldsymbol{X}_n \xrightarrow{m.s.} \boldsymbol{X}$ if each element of $\boldsymbol{X}_n$ converges to the corresponding element of $\boldsymbol{X}$. For a random vector to be asymptotically normal, we will write $\boldsymbol{X}_n \sim AN_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ if $\boldsymbol{\Sigma}_n$ has nonzero diagonal elements for all n sufficiently large, and if for every vector $\boldsymbol{h}$ such that $\boldsymbol{h}^T\boldsymbol{\Sigma}_n\boldsymbol{h} > 0$ for all n sufficiently large, then

$$\boldsymbol{h}^T\boldsymbol{X}_n \sim AN(\boldsymbol{h}^T\boldsymbol{\mu}_n, \boldsymbol{h}^T\boldsymbol{\Sigma}_n\boldsymbol{h}).$$

Thus a random vector is said to be asymptotically normal if all linear functions of it are asymptotically normal in the scalar sense.

Several times in the book (see Exercise T4.8 for example), we make use of the fact that a continuous function of an asymptotically normal sequence of random vectors is also asymptotically normal. We state this formally in the following theorem (see Serfling (1980), p. 122, for example).

Theorem A.6.6 Continuous Function Theorem.
Suppose that $\boldsymbol{X}_n \sim AN_r(\boldsymbol{\mu}, b_n\boldsymbol{\Sigma})$, *where* $\boldsymbol{\Sigma}$ *is a covariance matrix and* $b_n \to 0$ *as* $n \to \infty$. *Let*

$$\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_p(\boldsymbol{x}))^T, \quad \boldsymbol{x} = (x_1, \ldots, x_p)^T,$$

be a vector-values function for which each component function $g_i(\boldsymbol{x})$ *is real-valued and totally differentiable. Let* $\boldsymbol{G}$ *be the* $p \times r$ *matrix whose* $(i,j)^{th}$ *element is the partial derivative of* g_i *with respect to* x_j, *evaluated at the vector* $\boldsymbol{\mu}$. *Then*

$$\boldsymbol{g}(\boldsymbol{X}_n) \sim AN(\boldsymbol{g}(\boldsymbol{\mu}), b_n\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}^T).$$