1 Introduction

Our immediate environment is a magnificent tapestry of information-bearing signals of many kinds reaching us from many directions and from many sources, both large and small. Some signals arise as a response to some form of human-made illumination probing a scene or object of interest. Other signals occur naturally in the environment. Most signals are not readily compatible with our human senses. Only those signals such as optical signals in the visible band and certain acoustic signals are immediately compatible with our human senses. Other signals – such as electromagnetic signals in the infrared, ultraviolet, and X-ray bands, or waves in the radio and radar bands, or acoustic waves at ultrasonic frequencies – are not compatible with our natural senses, such as they are. To perceive any of these signals, which to our senses are invisible, sophisticated imaging algorithms are used to convert the sensed data into an understandable form such as an image.

A great variety of sensors now exist that collect signals and process these signals to form some kind of image, normally a visual image, of an object or a scene of objects. We refer to these as sensors for forming images, often requiring extensive computation to render the raw data into the form of an image. There are many kinds of sensors collected under this heading, differing in the size of the observed scene, as from classical microscopes to modern radio telescopes; in complexity, as from the simple lens to synthetic-aperture radars; and in the current state of development, as from photography to microscopy, holography, and tomography. Each of these systems collects raw sensor data and processes that sensor data into imagery that is useful to a user. This processing might be done by a digital computer, by an analog computer, or by an optical computer as may consist of a system of lenses. The development and description of a processing algorithm often requires a sophisticated theory and a precise mathematical formulation.

In this book, we bring together a number of signal-processing concepts that will form a background for the study and design of the many kinds of image formation system used for clinical evaluation or for remote surveillance. The signal-processing principles that we study include or adjoin the methods of medical imaging, classical radar and sonar systems, electromagnetic propagation, tomography, and physical optics, as well as estimation and detection theory.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

1.1 Image Formation

Mankind has designed a variety of devices that are used to observe the general environment and specific objects of interest in that environment. Images are formed by processing acoustic, pressure, or magnetic variations or by processing electromagnetic radiation in the radio and the microwave frequency bands, the infrared band, and the optical and X-ray bands. The many varieties of such medical imaging systems, and of radar and sonar systems, are examples of such systems. An image formation system may be active, using its own energy to illuminate the environment, or it may be passive, relying on signals already in the environment or signals produced by the scene itself.

The theory of image formation studies the design of signals to probe the environment as well as the design of computational procedures for the extraction of information from received signals within which that information may be deeply buried. As such, this theory comprises that branch of information theory that is explicitly concerned with the design of systems to observe the environment and with their performance of those systems. The theory herein is concerned specifically with the mathematical structure of the image-formation algorithms needed to extract information from the received signals.

An image formation system is any system that collects signals and creates an observable image by processing those signals by computation or otherwise to form that image. Figure 1.1 illustrates a computational image formation system partitioned into the "sensors" and the "algorithms." We will be concerned with the details of the image-formation algorithms and with the performance of those algorithms. We will be concerned with the physics of the sensors only insofar as is necessary to explain the relationship between the object of interest and the observed data or with the development of the algorithms.

The "image," which is the end product of the image formation system, is always some kind of depiction of an "actual" scene, usually a two-dimensional or three-dimensional scene, which we denote as $\rho(x, y)$ or $\rho(x, y, z)$. The scene may emit its own signals that the sensors intercept, or it may be probed with signals generated by the image formation system. Figure 1.2 shows a representative configuration in which the scene $\rho(x, y)$ is probed by a signal generated as a one-dimensional waveform. In this case, the sensors collect one or more reflected one-dimensional waveforms, $s_m(t)$, and from these reflected waveforms, the computational algorithms must form a suitable two-dimensional image of the scene. In this case, the computational task is to estimate the two-dimensional function, $\rho(x, y)$, (or a three-dimensional function, $\rho(x, y, z)$), when given a set of one-dimensional scattered waveforms, $s_m(t)$ for

Figure 1.1 A computational imaging system

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

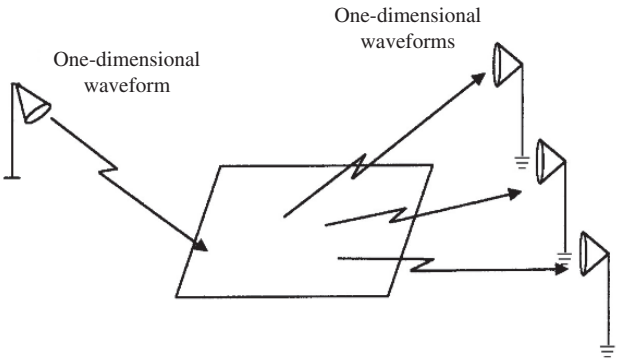

Figure 1.2 Probing a scene with waveforms

$m = 1, \ldots, M$, that depend on $\rho(x, y)$. The task of forming an image of an object or scene from a relevant signal is called an *inverse problem*. Among the most useful mathematical tools that we will develop for this task are the two-dimensional Fourier transform, the projection-slice theorem, and the ambiguity function. Probability theory, especially the notion of the likelihood function, is also an important tool in later chapters.

The many forms of reconstructed imagery, such as medical imagery and radar imagery, may often look very different from the visual imagery or conventional photographs of that same scene or object. This means that the user of that sensor may need training and experience in interpreting the generated image. To the novice, it may seem to be a limitation of that specialized image, but a more sophisticated view is that a new sensor opens a new window in our way of perceiving reality. A bat or a dolphin lives in a world that is perceived in large measure by means of acoustic or sonar data. This kind of sensor has nothing like the high angular resolution of our optical world, yet it does have other attributes, such as a strong doppler shift and the ability to resolve objects instantly by their velocities. Because it uses a different kind of data, the dolphin or the bat undoubtedly perceives the world differently from the way in which we do. Thus, it may be argued that modern image formation does change the way that society sees the world around us.

One way of defining the kind of imaging system to be studied is as a system in which raw signals in the environment that the human cannot sense directly are turned into processed signals that are compatible with one of the human senses. Thus a radar receiver converts an electromagnetic wave into a visual image compatible with the human eye, and a tomographic medical scanner turns an X-ray signal into a visual image of an anatomy. The image need not be a realistic replica of a photograph. Other details may be more important. An X-ray image of the human body does not look like a photograph of a human skeleton, but it may be preferred by the diagnostician because it contains useful information of other kinds.

An important topic also studied herein is the relationship between the illumination at the input of an aperture, such as an antenna or a lens, and the wavefront radiated

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

by that illumination. The reflection of these wavefronts, however, will be modeled in a simple way. Simplified models for both reflection and spontaneous emission will be adequate for most of our purposes. The detailed relationship between the wave incident on a reflecting object or scene and the wave reflected by that object or scene is called the *forward problem*. The forward problem is of interest in this book only insofar as it sets up the inverse problem. This is the problem of forming an image of that object or source from the reflected signal or from other observed data.

Radar and sonar are surveillance systems that are included among the imaging systems that will be studied. Originally, radar and sonar systems used simple waveforms and simple processing techniques that could be implemented with simple electronic circuits such as filters and threshold detectors. But over the years a new level of sophistication began to find its way into many of these systems. By maintaining a precise phase record of long-duration signals, and processing the signals phase-coherently, one can obtain new levels of system performance. Systems that depend on phase coherence over time are called *coherent surveillance* systems. Some early coherent systems in the radar bands were designed to use optical processing. More recently, digital processing of coherent electromagnetic waveforms has become practical.

Imaging algorithms depend on the angles available for viewing the target. Most radar systems view a target from only a single, or limited, viewing angle. The same is usually true for microscopy, photography, and astronomy. Medical imaging, in contrast, is often able to view an object from many angles, and the principles of tomography can be applied. Our goal is to develop a general theory of imaging systems in a common mathematical setting. We will be concerned with a range of processing algorithms, such as those used for forming the images of remote radar reflectors, X-ray or magnetic-resonance tomography, microscopy, and astronomy.

1.2 The History of Image Formation

The subject of image formation consists of the common overlap of a number of well-developed subjects such as physical optics, electromagnetics, and signal processing. From a broad point of view, the historical roots of imaging go back to the roots of these various subjects. We are interested here in a narrower view of this history, especially the history of clinical imaging systems, diffraction imaging systems, surveillance systems, and tomography. Our brief discussion in this section serves only to sketch the historical background of the material in this book.

Many kinds of imaging systems were developed independently, but share common fundamentals of signal processing and a common mathematical framework. These include: optical imaging, holography, medical imaging, radio astronomy, sonar beamforming, microscopy, diffraction crystallography, imaging radars, moving-target detection radars, as well as more recent topics such as seismic processing and passive source location.

Optical image formation systems within the topic of photography are among the earliest, the most developed, and the most familiar to the user. Credit for the invention

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

of photography is usually given to Niépce, who produced a photograph in 1826 or 1827 using the three principal components: an aperture, a lens, and a photosensitive medium. Photography passed into common use in 1839 after the work of Daguerre. Photographic imaging systems may be passive, using reflected light and occasionally using radiated light, or they may be active, using a source of light to illuminate the scene. The optics of a basic photographic system is adequately described using geometrical optics, but modern, high-performance multitone photography is based on wave optics. Multitone optical images are usually quite sharp with high resolution and excellent color contrast.

Image formation systems may use passive radiation in the infrared bands. These systems are similar to optical systems, but can form images of temperature variations in a scene because the intensity and wavelength of the radiation emitted by an object varies according to the temperature of that object.

Imaging was introduced into medical diagnostics by Roentgen in 1895 with the invention of X-ray radiography, which exposes a photographic film to X-rays transmitted through a body, then processes that film. Edison, by introducing an X-ray-sensitive fluorescent screen or fluoroscope in 1896, eliminated the delay required to process the film. The development of X-ray tomography in the modern sense of computerized image reconstruction for medical applications began in Great Britain. The key feature, based on the projection-slice theorem and the Radon transform, is the algorithmic reconstruction of images from their X-ray projections, first developed by Cormack in 1963 and reduced to practice by Hounsfield in 1971. The 1979 Nobel prize in physiology and medicine was awarded to Hounsfield and Cormack for the development of computerized tomography. The ideas of tomography are closely related to similar methods used in radio astronomy, especially the formulation of reconstruction algorithms by Bracewell (1956). Other kinds of computerized tomography are now in use for medical diagnostic imaging systems. In addition to the method of projection tomography based on X-ray projections, there are the methods of emission tomography and diffraction tomography. Photon- or positron-emission tomography (PET) based on radioisotope decay was proposed by Kuhl and Edwards (1963).

Magnetic resonance imaging (MRI) is yet another kind of tomographic imaging system based on magnetic excitation of atomic nuclei and the induced magnetization of the hydrogen nuclei distribution. The ground-breaking idea that enables MRI – for which Lauterbur and Mansfield shared the 2003 Nobel prize in medicine – is to use gradients of a magnetic field to encode spatial information into the transient response of a nuclear spin system after excitation by a magnetic pulse. Whereas X-ray tomography gives an image of the electron density, MRI gives an image of the distribution of hydrogen nuclei (isolated protons) in a body, though in principle it can be tuned to observe instead the distribution of other species of nuclei. The physical phenomenon of nuclear magnetic resonance had been observed independently in 1946 by Bloch and Purcell, for which they received the 1952 Nobel prize in physics. It was later realized that magnetic resonance effects varied with the kind of tissue excited but it was not known how to use this effect to make images. Lauterbur conceived and demonstrated his method of using the magnetic resonance phenomenon to form images by spatially

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

encoding the magnetic field, for which Lauterbur shared in the 2003 Nobel prize in medicine. Since then, magnetic resonance imaging has become an important modality in medical diagnosis. By using both static and time-varying magnetic excitation fields, an MRI system causes all nuclei of a given kind – selected by the resonance frequency of those nuclei – to precess (or oscillate), but with an amplitude and frequency modulation that depends on position as determined by the magnetic excitation at that position. The magnetization that is produced by the selected species of nuclei, usually hydrogen nuclei, is measured and sorted by frequency analysis. Because of the spatially-varying magnetic excitation, the frequency distribution of the induced magnetic field corresponds to the spatial distribution of the sources of radiation, which equates to the spatial density distribution of the target nuclei. Repeated scans at different angles allows the methods of tomography to be used to sort the data in other ways. Sophisticated mathematical algorithms based on the methods of tomography have been developed to so extract a high-resolution image of the hydrogen density distribution from the frequency distribution for each of multiple projections of measured magnetic resonance data.

The diffraction of X-rays by crystals was demonstrated in 1912 by Max von Laue, thereby demonstrating the wave properties of X-rays. Sir William Henry Bragg then immediately inverted the point of view to turn this diffraction phenomenon into a way of probing crystals, which has since evolved into a sophisticated imaging technique. The 1914 Nobel prize in physics was awarded to von Laue, and the 1915 Nobel prize in physics was awarded to Bragg and his son, Sir William Lawrence Bragg, who formulated the famous Bragg law of diffraction. This early work was directed toward finding the lattice structure of the crystal as a whole, but was not much concerned with the structure of the individual molecules making up the crystal. Attention soon turned to the finer question of finding the scattering structure within an individual cell of the crystal, and so to form an image of the molecule by processing that signal. A difficulty of this task is that, because of the small wavelength of X-rays, the phase of the diffracted X-ray wavefronts cannot be measured. Only the intensity (or amplitude) can be measured. Herbert Hauptman and Jerome Karle (1953) showed how to bypass this problem of missing phase by using prior knowledge about the molecules that compose the crystal, for which they shared the 1985 Nobel prize in chemistry. Earlier, in 1953, James Watson and Francis Crick – using the X-ray diffraction images produced by Rosalind Franklin – discovered the structure of the DNA molecule, for which they shared the 1962 Nobel prize in medicine.

Closely related to the methods of the Fourier transform and signal processing are many kinds of optical processing, many of them using diffraction phenomena that are describable in terms of the two-dimensional Fourier transform. A method known as the *schlieren method* was proposed by Jean Foucault in 1858 as a way to image density variations of air. Fritz Zernicke in 1935 developed phase-contrast methods to improve microscopy images, for which he was awarded the 1953 Nobel prize in physics. Aaron Klug developed methods for the imaging of viruses using the diffraction of electron microscope images, for which he won the 1982 Nobel prize in chemistry.

The 1914 Nobel prize in chemistry was awarded to Betzig, Moerner, and Hell for the development of superresolution fluorescence microscopy.

Dennis Gabor, influenced by the techniques used in microscopy and crystallography, proposed the idea of holography in a series of papers in 1948, 1949, and 1951. He originally intended holography as a method of microscopy but later as a replacement for photography. The work earned Gabor the 1971 Nobel prize in physics. Gabor realized that, whereas conventional photography first processes the raw optical wavefront to form an image which is then recorded on film, it is also possible to record the raw optical wavefront on the photographic film directly and place the processing of the optical wavefront in the future with the viewer. He called his method for the photographic recording of the raw optical data a *hologram*. Because the raw optical data contains more information than a final photographic image, in principle, the hologram can be used to create images superior to a photograph. Most striking in this regard is the creation of three-dimensional images from a two-dimensional hologram. Holography is technically much more difficult than photography because recording the raw optical data requires precision on the order of the optical wavelengths. For this reason, the idea of holography did not immediately draw the attention it deserved. Holography became more attractive after the invention of the laser and also after the more practical reformulation of the method by Leith and Upatnieks (1962), which was strongly influenced by Leith's other work on the optical processing used in synthetic-aperture radar.

Early radars used simple electronic circuits for processing the received signal while modern radars may use processing that is quite sophisticated. Sophisticated radar signal processing first appeared in the development of those imaging radars known as synthetic-aperture radars. The principle of such radars has been credited to a suggestion in 1951 by Wiley, although he did not then publish his ideas nor did those ideas then result directly in the construction of such a radar. Wiley observed that, whereas the azimuthal resolution of a conventional radar is limited by the width of the antenna beam, each reflecting element within the antenna beam from a moving radar has a doppler frequency shift that depends on the angle between the velocity vector of the radar and the direction to the reflecting element. Thus he concluded that a precise frequency analysis of the radar reflections would provide finer along-track resolution than the azimuthal resolution defined by the antenna beamwidth. The following year, a group at the University of Illinois arrived at the same idea independently, based upon frequency analysis of experimental radar returns. During the summer of 1953, these ideas were reviewed by the members of a summer study, "Project Wolverine," at the University of Michigan and plans were laid for the development of synthetic-aperture radar. It was recognized that the processing requirements placed extreme demands on the technology of the day. Many kinds of analog processors (filter banks, storage tubes, etc.) were tried. Meanwhile, Emmett Leith, at the University of Michigan, turned to the processing ideas of holography and adapted the optical processing techniques to satisfy the processing requirements for radar. In 1957, by using optical processing, the first synthetic-aperture radar was successfully demonstrated. Later, Green (1962) proposed

the use of the range-doppler techniques of synthetic-aperture radar for remote radar imaging of the surface of rotating planetary objects. This method is closely related to synthetic-aperture radar except using the rotation of the object itself to provide relative motion. High-resolution radar images of Venus from the earth gave us our first view of the surface of that planet unobstructed by the cloud cover of that planet.

Optical processors[1] are analog processors using the Fourier transforming property of a lens to compute two-dimensional Fourier transforms. Early on, these have been the processors of choice for imaging radars because of the sheer volume of data that can be handled. However, to form an image with an optical processor requires developing the photographic film twice within the processing, once to create the optical input signal and once to record the output. Optical processors are very sensitive to vibration, and so they are limited by the environment and also by the form of computations that can be included. Hence attention now has turned to other methods for processing. The advent of high-speed, digital array processors has had a large impact on the massive processing needed for synthetic-aperture imagery, and optical processing now plays a diminished or vanishing role.

The development of search radars for the detection of moving targets is spread more broadly, and individual contributions are not as easy to identify. From the first use of radar, it was recognized that the need to detect moving targets could be satisfied by using the doppler shift on the return signal. A moving object causes a doppler-shifted echo. However, the magnitude of the doppler shift is only a very small fraction of the transmitted pulse bandwidth. At that time, the technology did not exist to filter a faint, doppler-shifted signal from a strong background of signals echoed from other stationary emitters. Hence the development of search radars did not depend so much on invention at the conceptual level as it did on the development of technology to support widely understood requirements. By the end of World War II, radars had been developed that used doppler filters to suppress the clutter signal reflected from the stationary background. These early radars used simple delay lines to cancel the stationary return from one pulse with the (nearly identical) return from the previous pulse, thereby rejecting signals with zero doppler shift. In this way, large rapidly moving objects could be detected from stationary radar platforms and the radial velocity of these objects could be estimated.

Later, the requirements for search radars shifted to include moving, airborne radars for observing small, slowly moving target objects at long range. It then became necessary to employ much more delicate techniques for finding a signal return within a large clutter background. These techniques employ coherent processing with the aid of large digital computers.

Like radar, sonar is based on the reflection of a passband waveform from an object or scene, in the case of sonar, it is an acoustic wave. The carrier frequency might typically be between one kilohertz and one megahertz. Although sonar is similar to radar in principle, the speed of propagation is smaller by a factor of approximately 10^6, leading to practical consequences in beamforming. The concept of a synthetic-aperture radar

[1] Not to be confused with photonic processors.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

leads naturally to the notion of a synthetic-aperture sonar. High-resolution synthetic-aperture sonar using hydrophone arrays for beam steering has been developed for imaging the ocean floor and for other applications.

Meanwhile, astronomers had come to realize that a large amount of astronomical information reaches the earth in the microwave bands. Astronomers are well grounded in optical theory where beamwidths smaller than one arc second are obtained. In the microwave band, a comparable beamwidth requires a reception antenna that is many miles in diameter. Under the impact of wind, ice, and temperature gradients, such an antenna would need to be mechanically rigid. Clearly, such antennas are not practical. Around 1952, Martin Ryle, at the University of Cambridge, began to study methods for artificially creating a kind of an aperture by pairwise combinations of many individual antenna elements, or by allowing the earth's rotation to sweep an array of fixed antenna elements through space. In retrospect, this development of radio astronomy may be viewed as a passive counterpart to the development of active synthetic-aperture radar. The aperture is synthesized by recording the radio signal received at two or more antenna elements and later processing these records coherently pairwise within a digital computer. The first such radio telescope was the Cambridge One-Mile Radio telescope completed in 1964, followed by the Cambridge Five-Kilometer radio telescope in 1971. More recently, other synthetic-aperture radio telescopes have been built and put into operation throughout the world. (The continent-sized Very Large Baseline Array has an angular resolution of 0.0002 arc second.) For the development of synthetic-aperture radio telescopes, Ryle was awarded the 1974 Nobel prize in physics (jointly with Hewish who discovered pulsars with the radio telescope). Much of our knowledge of the extragalactic universe comes from the signal-processing algorithms that form the galactic images from the data gathered by the radio telescope antennas.

1.3 Baseband and Passband Waveforms

We will have frequent occasion to use real or complex baseband signals and also occasions to use passband signals. A real *baseband signal*, $s(t)$, is any real function of time with its spectral energy density concentrated near zero frequency. The baseband signal $s(t)$ may also be called a *baseband waveform* when it is regarded as a complicated signal or a *baseband pulse* when it is regarded as a relatively simple signal of finite energy. The *support* of $s(t)$ is the closure of the set of t for which $s(t)$ is nonzero.

A *complex baseband signal*, $s(t) = s_R(t) + js_I(t)$, where $j = \sqrt{-1}$, is any complex function of time with its spectral energy density concentrated near zero frequency. The *real* (or *in-phase*) component $s_R(t)$ and the imaginary (or *quadrature*) component $s_I(t)$ are both real baseband signals. The complex baseband signal $s(t)$ may also be called a *complex baseband waveform* or a *complex baseband pulse* as may be appropriate.

A *passband signal*, which is denoted by $\widetilde{s}(t)$, with a tilde overbar is a function of the form

$$\widetilde{s}(t) = s_R(t)\cos 2\pi f_0 t + s_I(t)\sin 2\pi f_0 t,$$

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

where f_0 is a constant known as the *carrier frequency* and $s_R(t)$ and $s_I(t)$ are real functions of time whose Fourier spectra $S_R(f)$ and $S_I(f)$ are zero for $|f| \geq f_0$. The signals $s_R(t)$ and $s_I(t)$ are called the *modulation components* of $\widetilde{s}(t)$. For a radar system, the carrier frequency f_0 lies somewhere in the interval from 0.1 to 35 gigahertz and is often in the interval from 1 to 10 gigahertz. For a sonar system, f_0 is usually measured in kilohertz. For an ultrasound system, f_0 may be measured in megahertz.

The passband signal $\widetilde{s}(t)$ may also be called a *passband waveform*, usually when it is regarded as a complicated signal; or a *passband pulse*, such as when it is regarded as a relatively simple signal of finite energy.

The *complex baseband signal* $s(t)$ corresponding to the passband signal $\widetilde{s}(t)$ is

$$s(t) = s_R(t) + js_I(t).$$

The real passband signal $\widetilde{s}(t)$ corresponding to the complex baseband signal $s(t)$ is[2]

$$\widetilde{s}(t) = \mathrm{Re}[s(t)\mathrm{e}^{-j2\pi f_0 t}].$$

The signals $\widetilde{s}(t)$ and $s(t)$ are regarded as essentially the same signal but for the detail of the multiplying complex exponential. To emphasize this, these may be called the *real passband representation* and the *complex baseband representation* of the same signal. It is often convenient to suppress the real part operator and write

$$\widetilde{s}(t) = s(t)\mathrm{e}^{-j2\pi f_0 t}.$$

In such a case, this is called the *complex passband representation* of the signal.

There are two reasons for replacing the passband signal $\widetilde{s}(t)$ with the complex baseband signal $s(t)$. From the notational point of view, the complex baseband signal is preferred because the complex baseband signal is notationally more compact than the passband signal, and mathematical manipulations of complex baseband equations exactly mimic mathematical manipulations of the corresponding passband equations and are much easier. Moreover, within a transmitter or receiver, it is often convenient to translate a real passband signal into the complex baseband representation. Ultimately, the simplest and most rewarding point of view is to think of the complex baseband signal as the more fundamental form which is temporarily represented as a passband signal for purposes of transmission and reception. While we study it and process it, the signal is a complex baseband signal; when we transmit it and receive it, the signal is a passband signal. To convert between the two forms is trivial, and is often the last operation in a transmitter and the first operation in a receiver.

[2] The sign in the exponent is arbitrary. It is chosen here so that Fourier transform relationships in optics and antenna theory have the conventional form. This choice leads to the positive sign convention appearing in the passband waveform. However, the opposite sign convention is used in modulation theory.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

1.4 Monodirectional Waves

Many imaging modalities are based on the processing of waves that are reflected from objects in a scene. The image formation algorithms are developed based on the structure and behavior of waves. A wavefront propagating in free space may have a complicated structure, both temporally and spatially. To gain an understanding of the general case, one can begin with a study of the simple case of a *plane wave*. More complicated situations can be built up from multiple plane waves. The Huygens–Fresnel principle, which is developed in Chapter 4, describes how any planar surface in free space crossed by a wave can be viewed as the source of that wave.

Physically, a wave may be a time-varying and space-varying electric or magnetic vector field associated with an electromagnetic wave, or it may be the time-varying and space-varying pressure field associated with an acoustic wave. A wave may be a vector function as in the case of the electromagnetic wave, or it may be a scalar function, as in the case of the acoustic wave. Our primary concern is with the mathematical description of the wave. Usually, we are content to deal with scalar-valued waves because of analytical simplicity. Although an electromagnetic wave is a vector-valued wave, this property of the wave does not often affect the properties of propagation that are of interest herein. With some exceptions, the wave can be regarded as a scalar wave for most of our purposes.

The propagation of electromagnetic waves at optical frequencies obeys the same fundamental laws as it does at microwave frequencies. However, the great difference in the wavelengths leads to a difference in the phenomena that we perceive. The wavelength of a microwave is on the order of centimeters, while the wavelength of a light wave is on the order of a micron. A microwave antenna rarely has dimensions of more than a few hundred wavelengths – and usually much less – while an optical lens has dimensions of more than 10^4 wavelengths. Consequently, an everyday optical beam is usually much sharper than a microwave beam and often is described adequately by geometrical optics and ray tracing.

Monochromatic Monodirectional Waves

A *monochromatic* wave is a wave at a single frequency. A *monodirectional wave* is a wave traveling in a single direction. Mathematically, a spatially uniform, monodirectional, monochromatic, scalar plane wave traveling in the z direction is given by

$$\begin{aligned}\widetilde{s}(t,x,y,z) &= A\cos(2\pi f_0(t-z/c)+\theta)\\ &= A\cos(2\pi f_0 t - kz + \theta),\end{aligned}$$

where the constant $k = 2\pi f_0/c = 2\pi/\lambda$ is called the *wave number* and λ is called the *wavelength*. This passband wave is also written as

$$\begin{aligned}\widetilde{s}(t,x,y,z) &= \mathrm{Re}[Ae^{-j\theta}e^{-j\left(2\pi f_0(t-z/c)\right)}]\\ &= \mathrm{Re}[Ae^{-j\theta}e^{-j\left(2\pi f_0(t-kz)\right)}],\end{aligned}$$

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

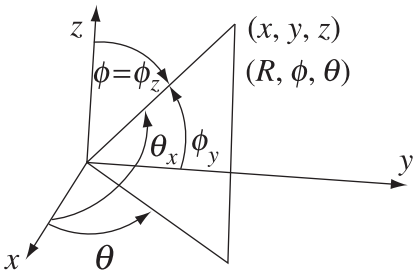

Figure 1.3 Direction cosines and spherical coordinates

where $Ae^{-j\theta}$ is called the *complex amplitude* of the wave at $z = 0$. The *complex baseband representation* of this wave at arbitrary z is

$$\begin{aligned} s(x,y,z) &= Ae^{-j\theta}e^{j2\pi f_0 z/c} \\ &= Ae^{-j\theta}e^{jkz}, \end{aligned}$$

with the time dependence now removed. This is the complex baseband representation of a monodirectional, monochromatic wave moving in the z direction. Such a wave is called a *plane wave*. A plane wave has the same value at every point of the wavefront plane.

The most general form of a spatially uniform, monodirectional, monochromatic wave that satisfies the wave equation is given by

$$\begin{aligned} \tilde{s}(t,x,y,z) &= A\cos(2\pi f_0(t-(\alpha x+\beta y+\gamma z)/c)+\theta) \\ &= A\cos\left(2\pi f_0 t-\alpha kx-\beta ky-\gamma kz+\theta\right). \end{aligned}$$

The variables α, β, and γ are called *direction cosines*. The direction cosines specify the direction of travel of the plane wave. They are equal, respectively, to the cosines of the angles between the direction of travel of the plane wave and the three coordinate axes:

$$\begin{aligned} \alpha &= \cos\phi_x, \\ \beta &= \cos\phi_y, \\ \gamma &= \cos\phi_z. \end{aligned}$$

The direction cosines are related to spherical coordinates, as shown in Figure 1.3, by

$$\begin{aligned} \alpha &= \cos\theta\sin\phi = \cos\phi_x, \\ \beta &= \sin\theta\sin\phi = \cos\phi_y, \\ \gamma &= \cos\phi = \cos\phi_z, \end{aligned}$$

and so they are related by

$$\alpha^2+\beta^2+\gamma^2 = 1.$$

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

The angles θ and ϕ are two of the three angles, called eulerian angles, relating two coordinate systems.

Three alternative complex baseband representations of the monodirectional, monochromatic wave with direction cosines α, β, and γ are[3]

$$\begin{aligned} s(x,y,z) &= A\mathrm{e}^{-j\theta}\mathrm{e}^{j2\pi f_0(\alpha x+\beta y+\gamma z)/c} \\ &= A\mathrm{e}^{-j\theta}\mathrm{e}^{jk(\alpha x+\beta y+\gamma z)} \\ &= A\mathrm{e}^{-j\theta}\mathrm{e}^{j(k_1 x+k_2 y+k_3 z)}. \end{aligned}$$

The quantities k_1, k_2, and k_3 are called the *wave numbers* of the plane wave. The wave numbers are related to the direction cosines by

$$\begin{aligned} k_1 &= (2\pi f_0/c)\alpha = (2\pi/\lambda)\alpha, \\ k_2 &= (2\pi f_0/c)\beta = (2\pi/\lambda)\beta, \\ k_3 &= (2\pi f_0/c)\gamma = (2\pi/\lambda)\gamma. \end{aligned}$$

The vector $\boldsymbol{k} = (k_1, k_2, k_3)$ is called the *vector wave number* of the plane wave.

The complex baseband representation using complex exponentials is more convenient to work with than is the passband representation. The real passband representation is recovered by

$$\widetilde{s}(x,y,z,t) = \mathrm{Re}\big[s(x,y,z)\mathrm{e}^{-j2\pi f_0 t}\big].$$

Time-Varying Monodirectional Waves

When the complex amplitude $A\mathrm{e}^{-j\theta}$ is replaced by a *time-varying* complex amplitude $A(t)\mathrm{e}^{-j\theta(t)}$, the waveform is no longer monochromatic. A monodirectional waveform with time-varying amplitude and phase traveling in direction (α, β, γ) has the general form

$$\widetilde{s}(x,y,z,t) = A(t-\tau(x,y,z))\cos\big(2\pi f_0(t-\tau(x,y,z)) + \theta(t-\tau(x,y,z))\big),$$

where $\tau(x,y,z) = (\alpha x + \beta y + \gamma z)/c$. This is called the *passband representation* of the time-varying monodirectional wavefront. Using the complex amplitude $A(t)\mathrm{e}^{-j\theta(t)}$, the passband representation can be written concisely as

[3] A wave of the form

$$\widetilde{s}(t,x,y,z) = A(x,y)\cos(2\pi f_0(t-z/c)+\theta),$$

does not satisfy the wave equation

$$\frac{\partial^2\widetilde{s}}{\partial x^2} + \frac{\partial^2\widetilde{s}}{\partial y^2} + \frac{\partial^2\widetilde{s}}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2\widetilde{s}}{\partial t^2}$$

whenever $A(x,y)$ is not the constant A. Consequently, it is not properly among the waves that are studied herein. Sometimes, it may be convenient to write a wave in this form. In such cases, $\widetilde{s}(t,x,y,z)$ should be regarded only as an approximation (geometrical optics) of a wave that does satisfy the wave equation. This approximation is studied in Chapter 4.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

$$\begin{aligned}\widetilde{s}(x,y,z,t) &= \mathrm{Re}\big[A(t-\tau(x,y,z))\mathrm{e}^{-j\theta(t-\tau(x,y,z))}\mathrm{e}^{-j2\pi f_0(t-\tau(x,y,z))}\big] \\ &= \mathrm{Re}\big[s(t,x,y,z)\mathrm{e}^{-j2\pi f_0 t}\big],\end{aligned}$$

where $s(x,y,z,t)$ is the complex baseband representation of the wavefront given by

$$s(x,y,z,t) = A(t-(\alpha x+\beta y+\gamma z)/c)\mathrm{e}^{-j\theta(t-(\alpha x+\beta y+\gamma z)/c)}\mathrm{e}^{-j2\pi f_0(t-(\alpha x+\beta y+\gamma z)/c)}.$$

At the origin, the real passband waveform is

$$\widetilde{s}(0,0,0,t) = \mathrm{Re}\big[A(t)\mathrm{e}^{-j\theta(t)}\mathrm{e}^{-j2\pi f_0 t}\big].$$

Let $S(f)$ be the Fourier transform of $s(t,0,0,0)$. A narrowband wave is one for which the support of $S(f)$ is narrow compared to f_0 insofar as the needs of the application may require. In most of this book, narrowband waves are approximated as monochromatic waves having a single wavelength λ.

1.5 Wavefront Diffraction

A monochromatic monodirectional wavefront is a wavefront that is traveling in only one direction. A wavefront that is not monodirectional is the superposition of waves traveling in multiple directions. We will see that when the waveform amplitude is spatially varying in every plane, the waveform is no longer monodirectional. It is a superposition of such monodirectional plane waves traveling in multiple directions. The complex amplitude in the x,y plane of each plane wave now depends on the distribution or spectrum of the wavefront directions.

Space-Varying Monochromatic Waves

When multiple monochromatic waves are simultaneously traveling in a finite number of directions, indexed by ℓ, the composite wave at complex baseband is

$$s(x,y,z) = \sum_{\ell=1}^{L} A_\ell \mathrm{e}^{-j\theta_\ell}\mathrm{e}^{j2\pi f_0(\alpha_\ell x+\beta_\ell y+\gamma_\ell z)/c},$$

where α_ℓ, β_ℓ, and γ_ℓ are the direction cosines specifying the direction of the ℓth plane wave.

When monochromatic waves are simultaneously traveling in all directions with an infinitesimal amplitude in each direction, then the complex baseband representation of the wavefront becomes an integral over the extent of wavefront directions. A monochromatic wavefront that has a continuum of directions has the complex baseband representation

$$s(x,y,z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(\alpha,\beta)\mathrm{e}^{j2\pi f_0(\alpha x+\beta y+\gamma z)/c}\mathrm{d}\alpha\mathrm{d}\beta,$$

where the pair (α,β) of direction cosines specifies a direction and $\gamma = \sqrt{1-\alpha^2-\beta^2}$ is the third direction cosine. The term $a(\alpha,\beta)\mathrm{d}\alpha\mathrm{d}\beta$ is the infinitesimal complex amplitude of the wave propagation in direction (α,β). Even though the direction cosines

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

range only between -1 and 1, the limits of integration have been written from $-\infty$ to ∞. This allows some important flexibility later. For now, the excess region of integration can be temporarily suppressed by requiring that $a(\alpha, \beta) = 0$ when $\alpha^2 + \beta^2$ is larger than one. This constraint will be dropped later.

When the wave $s_0(x, y) = s(x, y, 0)$ in the plane $z = 0$ is specified, the integral

$$s_0(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(\alpha, \beta) \mathrm{e}^{j2\pi(\alpha x + \beta y)/\lambda} \mathrm{d}\alpha \mathrm{d}\beta$$

implicitly defines $a(\alpha, \beta)$ in terms of $s_0(x, y)$. This equation can be interpreted as an instance of the inverse two-dimensional Fourier transform. The function $a(\alpha, \beta)$ is called the *angular spectrum* of the input signal $s_0(x, y) = s(x, y, 0)$ in the plane $z = 0$. The angular spectrum completely describes the propagation of a monochromatic wave. The "input" in the plane at z equal to zero is $s_0(x, y)$, which implicitly determines the angular spectrum $a(\alpha, \beta)$. In turn, in the plane with z equal to d, the complex amplitude $s_d(x, y) = s(x, y, d)$ is given in terms of $a(\alpha, \beta)$ by the expression

$$s_d(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(\alpha, \beta) \mathrm{e}^{j(2\pi/\lambda)\sqrt{1-\alpha^2-\beta^2}} \mathrm{e}^{j2\pi(\alpha x + \beta y)/\lambda} \mathrm{d}\alpha \mathrm{d}\beta.$$

We follow this line of thought in Chapter 4 to derive the important Huygens–Fresnel principle.

Evanescent Waves

There is also a less-familiar, monochromatic and monodirectional solution of the wave equation called an *evanescent wave*. An evanescent wave is a wave of the form

$$\widetilde{s}(x, y, z, t) = \cos(2\pi f_0(t - (\alpha x + \beta y)/c)) \mathrm{e}^{-2\pi f_0 \gamma z/c},$$

satisfying the wave equation, where now the term involving z is a real decaying exponential. To satisfy the wave equation, (α, β, γ) must satisfy

$$\alpha^2 + \beta^2 - \gamma^2 = 1,$$

where here γ^2 is led by a negative sign.

An evanescent wave has the complex baseband representation

$$s(x, y, z) = \mathrm{e}^{j2\pi f_0(\alpha x + \beta y)/c} \mathrm{e}^{-2\pi f_0 \gamma z/c},$$

with γ now defined as

$$\gamma = \begin{cases} \sqrt{1 - \alpha^2 - \beta^2} & \text{for} \quad \alpha^2 + \beta^2 \leq 1, \\ \sqrt{\alpha^2 + \beta^2 - 1} & \text{for} \quad \alpha^2 + \beta^2 \geq 1. \end{cases}$$

The two lines are defined differently so that γ is real in both cases. Thus $\alpha^2 + \beta^2 - \gamma^2 = 1$ for evanescent waves.

The evanescent wave is an exponentially decreasing wave in the z direction. This wave is needed by the mathematics or the physics in order to meet boundary conditions that cannot be met with a propagating wave in the z direction.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

Clearly, the amplitude of the evanescent wave becomes infinite as z goes to negative infinity. Therefore an evanescent wave can only exist in a half-space and requires special boundary conditions on the boundary of this half-space. When the half-space is taken to be the half-space for which z is nonnegative, the boundary conditions are on the plane at which $z = 0$. The evanescent wave decays quickly with increasing z, becoming negligible after a few wavelengths. Along the plane at $z = 0$ in the direction specified by α and β runs an evanescent wave with a velocity of $c/\sqrt{\alpha^2 + \beta^2}$, which is smaller than c.

With the introduction of evanescent waves, the equation

$$s(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(\alpha, \beta) \mathrm{e}^{j2\pi f_0 z \sqrt{1-\alpha^2-\beta^2}/c} \mathrm{e}^{j2\pi f_0(\alpha x+\beta y)/c} \mathrm{d}\alpha \mathrm{d}\beta,$$

introduced earlier, can now be interpreted more generally. The infinite limits of integration allow the direction cosines to be larger than one. Physically, this allows evanescent waves to be included within the angular spectrum $a(\alpha, \beta)$.

Transverse Vector Waves

Besides scalar-valued waves, there are also vector-valued waves. A vector-valued wave may be regarded as three scalar-valued waves comprising the three components of the vector in a suitable coordinate system. A monodirectional, monochromatic vector wave at complex baseband has the form

$$\boldsymbol{s}(x, y, z) = [s_x \boldsymbol{i}_x + s_y \boldsymbol{i}_y + s_z \boldsymbol{i}_z] \mathrm{e}^{j2\pi f_0(\alpha x+\beta y+\gamma z)/c},$$

where $(\boldsymbol{i}_x, \boldsymbol{i}_y, \boldsymbol{i}_z)$ forms a triad of orthogonal unit vectors along the three axes of the coordinate system. When the three scalar components s_x, s_y, and s_z can be independently specified, then such a wave amounts to nothing more than three independent scalar waves.

There are certain vector waves of widespread physical interest satisfying a special constraint that makes the components dependent. These vector waves, called *transverse-vector waves*, are those waves that satisfy an additional constraint. Electromagnetic waves in free space are transverse vector waves.

A transverse vector wave is a vector wave that takes only values perpendicular to its direction of propagation. For the wave to be a transverse wave, the direction of the vector field must be perpendicular to the direction of propagation. The dot product of the field vector and the direction of propagation must be zero. For a plane wave, the direction of propagation $\alpha \boldsymbol{i}_x + \beta \boldsymbol{i}_y + \gamma \boldsymbol{i}_z$ is constant. The direction of the field is $s_x \boldsymbol{i}_x + s_y \boldsymbol{i}_y + s_z \boldsymbol{i}_z$. This means that for a transverse-vector plane wave, the dot product

$$s_x \alpha + s_y \beta + s_z \gamma = 0$$

must be satisfied as a side condition.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

1.6 Temporal and Spatial Coherence

The words "coherent" and "noncoherent" continually recur. These words are used to designate the quality of the phase angle of a passband waveform. Every passband signal is a sinusoid that can be expressed in terms of the time-varying amplitude and phase

$$\widetilde{s}(t) = A(t)\cos(2\pi f_0 t + \theta(t)),$$

where $A(t)$ is the time-varying amplitude, f_0 is the carrier frequency, and $\theta(t)$ is the time-varying phase. The complex baseband representation then has the form

$$s(t) = A(t)\mathrm{e}^{-j\theta(t)}.$$

The phase $\theta(t)$ may be intentional and known, or it may be partially or wholly unintentional and unknown. The phase angle may be random phase noise. When the phase angle $\theta(t)$ of the signal $\widetilde{s}(t)$ is known to the extent that knowledge of $\theta(t)$ is critical to the application, the signal $\widetilde{s}(t)$ is called a *coherent* signal. Otherwise, $\widetilde{s}(t)$ is called a *noncoherent* signal.

The term "coherent" may also arise in connection with the processing of a real passband signal, perhaps in the form of a complex baseband signal. The processing may be the kind known as *coherent processing*, which fully uses both $A(t)$ and $\theta(t)$, or the kind known as *noncoherent processing*, which makes only limited – or no – use of $\theta(t)$.

Coherence not only refers to a deterministic relationship between the phase angles of a waveform at different time instants, but may also refer to a deterministic relationship between the phase angle of two different waveforms, $\widetilde{s}_1(t)$ and $\widetilde{s}_2(t)$. The former case is then referred to as a *temporally* coherent waveform. The latter case is referred to as a *spatially* coherent waveform when a common wavefront is incident on two antennas or two lenses at different locations, or at two regions of the same antenna or lens. The deterministic relationship between points in a spatially coherent wavefront may be due to the different times at which the wavefront reaches those different points. Two signals, $\widetilde{s}_1(t)$ and $\widetilde{s}_2(t)$, may be spatially coherent even though they are jointly temporally noncoherent. For example, in photographic systems, the light from a point source incident on a lens may be temporally noncoherent, but across the lens it is spatially coherent. Otherwise, the lens could not focus the light into an image of that point source. Moreover, when there are multiple point sources, the light emitted by the multiple point sources can be mutually spatially noncoherent because the point sources are mutually noncoherent, yet the light reaching the lens from each individual point source can be spatially coherent.

A pulse train is a common example of a passband radar waveform. A pulse train has the form

$$\widetilde{p}(t) = \sum_{n=0}^{N-1} s(t - nT_r)\cos(2\pi f_0 t + \theta_0),$$

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

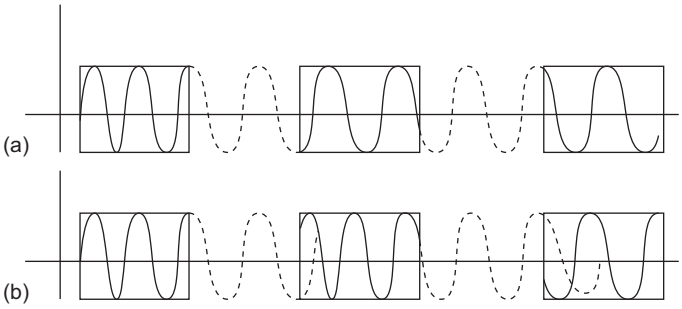

Figure 1.4 A coherent pulse train (a) and a noncoherent pulse train (b)

where $s(t)$ is a single pulse, T_r is a constant called the *pulse repetition interval*, and θ_0 is a constant. The pulse train consists of N uniformly spaced translates of the pulse $s(t)$ modulated onto the carrier $\cos(2\pi f_0 t + \theta_0)$. In complex baseband notation, the pulse train is denoted

$$p(t) = \sum_{n=0}^{N-1} s(t - nT_r)\mathrm{e}^{-j\theta_0}.$$

This pulse train might be called coherent to mean that θ_0 remains constant from pulse to pulse even though θ_0 may be unknown. Instead, the waveform might be called coherent to mean that the pulses all have the same known constant phase θ_0. Hence whether or not a given waveform is called coherent might depend on the circumstances of the discussion.

A pulse train in which the phase is not the same from pulse to pulse is given in the passband representation by

$$\tilde{p}(t) = \sum_{n=0}^{N-1} s(t - nT_r)\cos(2\pi f_0 t + \theta_n),$$

and in the complex baseband representation as

$$p(t) = \sum_{n=0}^{N-1} s(t - nT_r)\mathrm{e}^{-j\theta_n}.$$

This is called a noncoherent pulse train when the θ_n are random and independent (or weakly correlated). Then the θ_n may form a sequence of independent random variables, perhaps taking values uniformly between 0 and 2π. Figure 1.4 compares the coherent pulse train with a noncoherent pulse train. It is important to the usage here that the phase angles are unknown. When the phase angles are known, even though different, the waveform is a coherent waveform because the known values of the phase angles can be included in the processing of the waveform.

More generally, $\theta(t)$ may separate into two parts: a phase angle that is known, and a phase angle that is unknown. For an arbitrary waveform in the passband representation, this may be written as

$$\tilde{s}(t) = A(t)\cos[2\pi f_0 t + \theta_s(t) + \theta_n(t)].$$

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

The complex baseband representation is

$$s(t) = A(t)\mathrm{e}^{-j[\theta_s(t)+\theta_n(t)]},$$

where $\theta_s(t)$ is the intentional and known part of the phase modulation associated with the signal, and $\theta_n(t)$ is the unintentional and unknown part of the phase modulation. The unknown part is called *phase noise*. When described in this way, a coherent waveform could mean a waveform in which $\theta_n(t)$ is negligible, and a noncoherent waveform could mean a waveform in which $\theta_n(t)$ is not negligible. In some applications, surprisingly large values of unknown phase noise may be acceptable. Even phase errors as large as one radian can sometimes be tolerated, though with a significant loss of system performance. Chapter 15 is devoted to the quantitative analysis of the effect of phase error and phase noise on performance, including the notion of coherence in various situations.

1.7 Deterministic and Random Models

The usual goal of an image formation system is to form the best image of an object based on the available data. However, it is difficult to formulate a precise statement of optimality because a general criterion of optimality can be elusive and prior knowledge may be subjective. This is due partly to the fact that the underlying physical reality is much richer than the desired image, and it is difficult to state the real goal as an abstraction of the physical reality, and also because prior knowledge or prior assumptions about the image must be accommodated. Such considerations require that a model of the problem be developed. Such a model may be either deterministic or random. In the early chapters of this book, deterministic models of the image are usually used. Such models assume that a "true" image does exist, and our task is to estimate that image by processing the observed data. Randomness enters the problem in those chapters only because the measurements can be random or noisy. There is a single underlying image that is to be found.

In later chapters, we turn to a more abstract view of imaging, regarding the task as one of selecting an image from a space of possible images. In that more abstract view, an image is a realization of a random variable characterized by a probability distribution on a predefined space of images. The goal is not to pick the "true" image, but to select that image from the space of images that best explains the observed data. This reformulation of the task of imaging may be seen as nearly the same task as before, but it does suggest alternative approaches.

For these reasons, both to model measurement noise and to model a random image, probability theory inevitably enters the topics of this book. We will need the notions of a random variable and a random process. Here we briefly review some of the fundamentals of probability theory that are used.

The reason for introducing the topic of random variables described in this section and the topic of random processes described in more detail in Section 2.8 of Chapter 2 is to study randomness arising in various situations of imaging. Imaging theory uses

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

quantities such as probability density functions and correlation functions. They must be meaningful and known. Eventually, we even deal with the doubly vague situation in which the probability density function $p(x)$ associated with the random variable X is itself unknown. Such formulations provide structure and lead to useful procedures for image formation.

Random Variables

A *random variable*, X, consists of a set of values that the random variable can take and a probability distribution on this set of values. A random variable, X, may be restricted to a finite, or countable, number of values, in which case it is called a *discrete random variable*. Then it is characterized by the *probability vector* $\boldsymbol{p}$ with a finite, or countable, number of components denoted p_j. Likewise, a pair of discrete random variables, (X, Y), is associated with a joint probability distribution, $\boldsymbol{P}$, with an array of components denoted P_{jk}. A joint probability distribution is associated with *marginals*, defined by $p_j = \sum_k P_{jk}$ and $q_k = \sum_j P_{jk}$, and *conditionals*, defined by $Q_{k|j} = P_{jk}/p_j$ and $P_{j|k} = P_{jk}/q_k$. This leads to the *Bayes formula*

$$Q_{k|j} = \frac{q_k P_{j|k}}{\sum_k q_k P_{j|k}}$$

as a consequence of the definitions of marginals and conditionals.

A *real random variable* is a random variable that takes values in the set of real numbers. A real random variable may take values in a finite set of real numbers, in which case it is called a *discrete real random variable*, or values in a continuous set of real numbers, in which case it is called a *continuous real random variable*. Whereas a discrete random variable is described by a probability vector $\boldsymbol{p}$, a continuous random variable is described by a function, $p(x)$, called the *probability density function*, or a conditional function $p(x|y)$, called the *conditional probability density function*. We consider only discrete random variables and continuous random variables. We do not consider mixed random variables.

A discrete or continuous real random variable has a mean, $\bar{x}$, denoted by

$$\bar{x} = \sum_j p_j x_j,$$

or by

$$\bar{x} = \int_{-\infty}^{\infty} x p(x) \mathrm{d}x,$$

and a variance, σ^2, denoted by

$$\sigma^2 = \sum_j p_j (x_j - \bar{x})^2,$$

or by

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 p(x) \mathrm{d}x.$$

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

Similar definitions can be made for *complex random variables*, which are random variables taking values in the set of complex numbers.

An important random variable is a *gaussian random variable*, which is the only example of a random variable given in this section. Other random variables appear later in the book. The real gaussian random variable is defined by its probability density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\bar{x})^2/2\sigma^2}.$$

The gaussian random variable has the mean $\bar{x}$ and variance σ^2. Likewise, the complex gaussian random variable $X = X_R + jX_I$ with circular symmetry has probability density function

$$\begin{aligned} p(x_R, x_I) &= \frac{1}{2\pi\sigma^2} \mathrm{e}^{-|x-\bar{x}|^2/2\sigma^2} \\ &= \frac{1}{2\pi\sigma^2} \mathrm{e}^{-x_R^2/2\sigma^2} \mathrm{e}^{-x_I^2/2\sigma^2}, \end{aligned}$$

where $\sigma^2 = \mathrm{E}[X_R^2] = \mathrm{E}[X_I^2] = \mathrm{E}[XX^*]/2$ and $\mathrm{E}[X_R X_I] = 0$ for the circularly symmetric complex gaussian random variable.

A real (or complex) *multivariate random variable*, $\boldsymbol{X} = (X_1, \ldots, X_n)$, also called a *vector random variable*, with zero mean has a probability density function, $p(x_1, \ldots, x_n)$ and a covariance matrix, $\boldsymbol{\Sigma}$, whose ij entry is the expectation $\mathrm{E}[X_i X_j]$ (or $\mathrm{E}[X_i X_j^*]$). A covariance matrix is always nonnegative-definite because[4] $\boldsymbol{a}\boldsymbol{\Sigma}\boldsymbol{a}^\dagger = \boldsymbol{a}\mathrm{E}[\boldsymbol{X}\boldsymbol{X}^\dagger]\boldsymbol{a}^\dagger = \mathrm{E}[(\boldsymbol{a}\boldsymbol{X})^2]$, which is always nonnegative because it is the expectation of a squared term.

Random Processes

A *random process* or a *stochastic process*, $X(t)$, on the variable t consists of a set of functions that the random process can take and a probability distribution on this set of functions. The values that $X(t)$ can take may be continuous or discrete. The independent variable t can be continuous or discrete. Usually, a *discrete random process* refers to a random process for which t is discrete.

1.8 The Electromagnetic Spectrum

Signals throughout the electromagnetic spectrum are everywhere and carry a great deal of information, much of it hidden from our senses. Table 1.1 shows the remarkable twenty orders of magnitude of the electromagnetic spectrum that are of interest. The table shows the spectrum broken into bands annotated with individual names. The

[4] The symbol † denotes the transpose of a real-valued matrix or the complex conjugate of the transpose for a complex-valued matrix.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

Table 1.1 Common names of electromagnetic spectrum intervals

Radiation type	Frequency range (hertz)
Radio waves	2×10^4 to 1×10^9
Microwaves	1×10^9 to 3×10^{11}
Infrared	3×10^{11} to 4×10^{14}
Near infrared	1×10^{14} to 4×10^{14}
Visible	4×10^{14} to 7.5×10^{14}
Ultraviolet	1×10^{15} to 1×10^{17}
X-rays	1×10^{17} to 1×10^{20}
Gamma rays	1×10^{20} to 1×10^{24}

naming of these frequency bands is not standardized and the usage may vary somewhat. The boundary between the bands is not sharply defined. The various topics in this book range throughout the electromagnetic spectrum shown in Table 1.1.

The electromagnetic spectrum is labeled using one of three measurement units. These three measurement units are the frequency f, the wavelength λ, and the photon energy E, only one of which is mentioned in Table 1.1. These three quantities are related by the expressions

$$\lambda = c/f \qquad E = hf = h\lambda/c,$$

where c is a constant called the *speed of light* and h is a constant called the *Planck constant*. Each of these three measurement units is most convenient to use in a different region of the spectrum according to how the electromagnetic signal presents in that region. Only the frequency designation is given in Table 1.1.

Our senses lack the ability to observe most of the electromagnetic spectrum directly. Our sense of vision allows us to observe only the very narrow range of frequencies known as the visual band. Although our sense of touch is sensitive to frequencies in the near infrared and near ultraviolet, this sensitivity is only a vague awareness of radiation in these frequencies with no awareness of a corresponding image. The remainder of the immense electromagnetic spectrum, though teeming with information of many kinds, is outside of our immediate experience.

The electromagnetic spectrum is indeed full of signals. Both natural signals and man-made signals are present. Many image formation systems are based on sensing these electromagnetic signals throughout the electromagnetic spectrum. Even in the visible spectrum, our natural vision is now augmented by many man-made devices, such as eyeglasses, cameras, microscopes, and telescopes.

Table 1.1 partitions the electromagnetic spectrum according to general terms that are in common usage. A more systematic partition of a part of the spectrum is given in Table 1.2. The terms in this table are also in common use.

Other kinds of image formation systems are based on acoustic (seismic) signals. The acoustic spectrum is also large, extending into the ultrasound frequencies.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

Table 1.2 Decades of the electromagnetic spectrum

Decade designation		Frequency range
Very low frequency	(VLF)	3×10^0 to 3×10^1 kHz
Low frequency	(LF)	3×10^1 to 3×10^2 kHz
Medium frequency	(MF)	3×10^2 to 3×10^3 kHz
High frequency	(HF)	3×10^0 to 3×10^1 MHz
Very high frequency	(VHF)	3×10^1 to 3×10^2 MHz
Ultra high frequency	(UHF)	3×10^2 to 3×10^3 MHz
Super high frequency	(SHF)	3×10^0 to 3×10^1 GHz
Extremely high frequency	(EHF)	3×10^1 to 3×10^2 GHz

1.9 Imaging by Tomography

In many situations, such as in medical imaging, it is possible to observe projections of the two-dimensional object $\rho(x, y)$ (or the three-dimensional object $\rho(x, y, z)$) of a certain kind, even though direct observations of that object are not possible. Using X-rays, a nuclear beam, or magnetic resonance gradients, the object $\rho(x, y)$, as perceived by that energy source, can be integrated along lines. The results of these integrations are called *projections*. An image of $\rho(x, y)$ is computed from the set of its projections. The process of imaging from projections is known as *tomography*.[5] The central theorem of tomography is the *projection-slice theorem*. The set of all projections of $\rho(x, y)$ is called the *Radon transform* of $\rho(x, y)$. While $\rho(x, y)$ is not directly observable, the Radon transform of $\rho(x, y)$ is observable as a collection of projections.

A familiar example is an elementary X-ray projection of internal body organs. As a single ray passes along a line, say the y axis with x held constant and ignored for now, the intensity is attenuated at each y by an amount described by an attenuation function $\rho(y)$. That is, the intensity, denoted I', leaving a small interval of width Δy, centered at y_1, is related to the intensity, denoted I, entering that interval by

$$I' = I[1 - \rho(y_1)\Delta y].$$

This is approximated as

$$I' \approx I\mathrm{e}^{-\rho(y_1)\Delta y}$$

under the condition that the attenuation is small for a sufficiently small interval Δy. Over two consecutive intervals, each of width Δy, the intensity attenuation is described approximately as

$$I'' = I\mathrm{e}^{-\rho(y_1)\Delta y}\mathrm{e}^{-\rho(y_2)\Delta y}.$$

[5] The term tomography has been broadened, by some, to include other forms of medical imaging

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

Consequently, over a sequence of many such intervals, the output intensity I_{out} is related to the input intensity I_{in} by

$$I_{out} = I_{in} \mathrm{e}^{-\sum_i \rho(y_i)\Delta y}.$$

In the limit as Δy goes to zero,

$$\log_{\mathrm{e}} \frac{I_{in}}{I_{out}} = \int_{-\infty}^{\infty} \rho(y) \mathrm{d}y,$$

where the integration limits can be replaced by the support of the function $\rho(y)$.

By introducing a ray in the y direction at each value of x, passing through the two-dimensional function $\rho(x, y)$, one can define the *projection* $p(x)$ onto the x axis for each value of x:

$$p(x) = \int_{-\infty}^{\infty} \rho(x, y) \mathrm{d}y.$$

The function on the left is the projection of $\rho(x, y)$ onto the x axis.

In the general case, the attenuation of an X-ray at angle θ integrates the function $\rho(x, y)$ along each ray in the direction indicated by angle θ. The projection at angle θ

$$p_\theta(t) = \int_{-\infty}^{\infty} \rho(t\cos\theta - r\sin\theta, t\sin\theta + r\cos\theta) \mathrm{d}r,$$

at each t, consists of the integration of $s(x, y)$ along each ray in the r direction as a function of t. By varying the viewing angle θ, such projections can be observed from many directions. One wants to process such a set of projections to form an estimate of $\rho(x, y)$, as may show the internal organs of the body. The signal-processing topic of tomography is studied in Chapter 6. The central theorem of signal processing that underlies the methods of tomography is the projection-slice theorem, which is introduced in Chapter 3.

The origins of tomography can be traced back to 1917 when the Austrian mathematician Radon showed that the spatial function $\rho(x, y, z)$ can be reconstructed from the complete set of its projections. Because the reconstruction of images from projections arises in many diverse situations, it is not surprising that this mathematical principle, first discovered by Radon, was independently rediscovered many times and in many fields. It has been used in radio astronomy and in the field of electron microscopy. In the context of medical applications, tomography has led to important advances in the noninvasive imaging techniques available in recent years for clinical practice and medical research. Tomography is used in many other applications, such as geophysical applications, where it can be used for subsurface exploration, or in atmospheric sensing, where it can be used, for example, to form images of pollutant densities in the upper atmosphere.

In Chapter 6, the mathematical principles underlying tomography are studied, especially the projection-slice theorem, which relates the one-dimensional Fourier transform of the projection to the two-dimensional Fourier transform of the object. A number of algorithms for the reconstruction of images are described. The central idea of these algorithms is the method of back projection. Reconstruction of an arbitrary

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

object from projections is exact only when the uncountably infinite set of projections from all angles is known. However, mild prior conditions on the object, such as a spatially bandlimited Fourier transform, can soften this statement. Many good algorithms are known that compute an approximate reconstruction of an image from a finite set of its projections. These images may be satisfactory for practical applications even when the individual projections are weak or noisy.

In addition to projection tomography, many other important forms of tomography go by names such as *emission tomography*, *diffraction tomography*, *diffusion tomography*, and *coherence tomography*. These forms are also studied in Chapter 6. Emission tomography requires that the scene itself emits some form of emission that provides the received signal from which the image is computed. This usually means that the scene must receive an excitation that provides the energy for the emission. There are two methods that are in wide use to excite an object so that it will produce a useful signal. The important method of *magnetic resonance imaging* (MRI) uses a time-varying and spatially varying magnetic field to provide energy to the scene. This time-varying magnetic field causes isolated protons, or perhaps other selected nuclei, to resonate and thus generate signal-dependent magnetic fields. The magnetic field is intercepted, measured, and processed by the methods of tomography to form an image of the density of isolated protons (hydrogen atoms). Another method of excitation, called *positron-emission tomography*, uses a radioactive isotope that is selectively absorbed by a tissue of interest, usually a diseased tissue. The radioactive isotope then decays, thereby releasing radiation energy in the form of positrons. These positrons immediately combine with electrons to produce photons. The photons are captured by an array of photosensors. From the positions and times at which these photons are detected, an image is formed. Through this method, a specific tissue can be selectively imaged as a function of x and y by its tendency to acquire a particular radioactive isotope.

Diffraction tomography and diffusion tomography deal with situations in which the geometrical-optics approximation to propagation is not adequate. It may be necessary to treat wave propagation in a more exact way by considering the effect of diffraction. This is particularly important when observing details that are small compared to the relevant wavelengths. Another difficult instance of tomography is based on the propagation of a wave in a strongly scattering medium. This is the difficult topic of diffusion tomography.

A related form of tomography is *geophysical tomography* in which seismic waves are used to image geophysical features. Then the dispersion of the wave is not caused by diffraction, but rather is caused by scattering anomalies in the propagation medium.

1.10 Radar and Sonar Systems

A radar obtains information about an object or a scene by illuminating the object or the scene with electromagnetic waves, then processing the echo signal that is reflected from that object or scene and intercepted by the radar receiving antenna. A sonar obtains information about an object or a scene by illuminating the object or scene

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

with acoustic waves, then processing the echo signal that is reflected from that object and intercepted by the sonar hydrophones.

By using electromagnetic waves in the microwave bands, a radar is able to penetrate optically opaque media such as clouds, dust, soil, or foliage. In this way, it is possible to form radar images of objects hidden by such obstructions. Similarly, a sonar or ultrasound system can form an image of an object that is optically masked by an opaque medium.

Most instances of radar or sonar are *monostatic*. This means that the transmitter and the receiver are colocated. A monostatic radar may use the same antenna for transmission and reception, as is the usual case. For a *bistatic* radar or sonar, the transmitter and the receiver are at separate locations.

In some cases, the broad beamwidth of a radar antenna or a sonar hydrophone is appropriate as a way of viewing a large region of space. For reasons such as these, radar and sonar have long been popular as imaging systems for surveillance.

While there may be a great deal of difference between the propagation of electromagnetic waves and the propagation of acoustic or pressure waves, there is also a great deal of similarity.[6] This similarity carries over to radar and sonar systems. From our point of view, each is a system that forms a complex baseband pulse, $s(t)$, that is transmitted as the amplitude and phase modulation of the passband pulse $\widetilde{s}(t)$, and receives an echo pulse, $\widetilde{v}(t)$, that is a composite of delayed and frequency-shifted copies of the passband pulse $\widetilde{s}(t)$ and contaminated by noise. The transmitted pulse $\widetilde{s}(t)$ propagates at a velocity c over a path of length R_1 from the transmitter to the reflector, and then over a path of length R_2 from the reflector to the receiver. The received pulse $\widetilde{v}(t)$ is a superposition of echoes of the transmitted pulse from multiple reflectors. Because the received signal is contaminated by noise and other impairments, it is difficult to recognize individual reflectors. We are interested in methods of processing the received pulse to extract useful information from it. The same basic ideas apply equally to radar pulses and to sonar pulses although the propagation medium is not uniform for sonar. The terminology of the discussion will favor radar systems.

The received signal is distributed both in space across the aperture of an antenna and in time. The distribution in space may be processed by the antenna system to gather all of the received spatially distributed signal into a single, time-dependent signal. Simple linear processing of the signal across the aperture is usually summarized by referring to the shape and width of an antenna "beam." The time variations of the received signal are processed so as to determine the time-varying distance to the reflecting objects. The space distribution may be processed in other ways to determine the direction of arrival of the signal. The processing of the space distribution of the signal across an aperture is studied in Chapter 5. The processing of the time distribution of the signal is studied in Chapters 10, 11, and 12.

[6] As a transverse-vector wave, an electromagnetic wave also has the property of polarization which is sometimes useful to a radar system.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

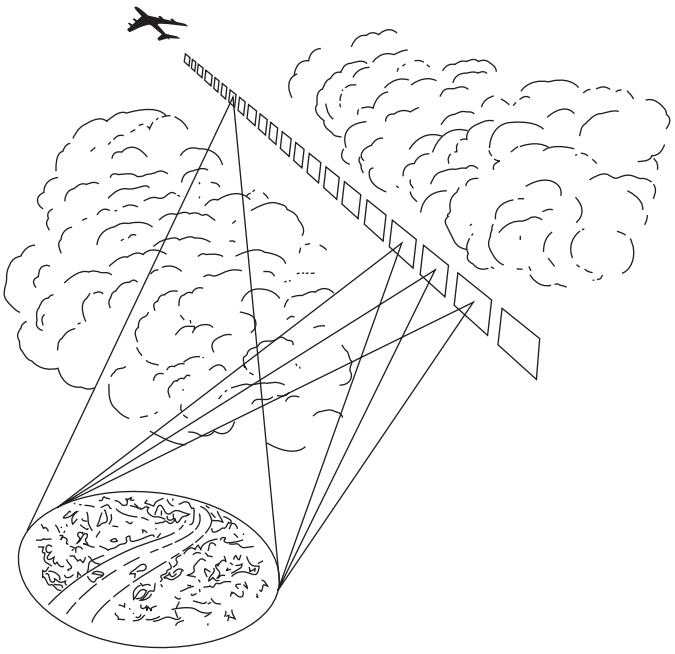

Figure 1.5 A spotlight-mode airborne radar

The typical airborne radar transmits a passband microwave signal consisting of a train of uniform pulses. A moving transmitter illuminates a scene with this waveform. This is illustrated in Figure 1.5, which shows the position of the antenna for each transmitted pulse. The figure shows a spotlight-mode radar in which the antenna or antenna system moves the antenna beam to illuminate a chosen scene. The pulses may be processed individually, in which case a noncoherent pulse train suffices. Such would be a rather simple radar for the detection of fixed or moving objects as described in Chapter 12.

A more advanced radar maintains coherence across the entire pulse train and processes the pulse train echo coherently as a whole. Coherent processing of a received passband signal is the form of processing that employs the carrier phase structure of the waveforms. A coherent system is informally defined as any system for which coherent processing is fundamental to its operation. The extraction of maximum information from a received passband signal requires a waveform that supports coherent processing over long time intervals, and this can lead to the use of sophisticated signal processing.

When the transmitter or receiver is in motion with respect to the object reflecting the signal, coherent processing becomes a potent technique because of the resulting frequency shifts, called *doppler*, in the echo. One system of this kind that is used for imaging is called a *synthetic-aperture radar* because of the heuristic notion of synthesizing a long, fixed antenna by the sequence of positions of a short, moving antenna as suggested by Figure 1.5. A synthetic-aperture imaging system depends on motion.

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

The motion can be described as a time-varying position. This description is often simplified as a straight line that is specified by an initial position and a constant velocity. Whenever this description suffices, the received electromagnetic signal depends on the scene parameters through the time delay and the doppler frequency shift of that signal.

The received echo signal corresponding to each pulse is converted to a precision optical or digital replica, maintaining both the amplitude modulation and the phase modulation. A history of such received pulses is accumulated, each pulse occurring at a slightly different position along the trajectory of the radar. From this history, an image of the illuminated scene is assembled by coherent processing.

The performance of a waveform for search or imaging is studied with the aid of a two-dimensional function called the *ambiguity function* or the *Woodward function*. The ambiguity function of any pulse $s(t)$ or pulse train $p(t)$ is unique for that pulse or waveform. The ambiguity function of a pulse or waveform is the key to understanding the performance of an imaging or detection radar that uses the pulse or waveform. The ambiguity function is studied in detail in Chapter 10.

Problems

1.1 Show that the gaussian density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\bar{x})^2/2\sigma^2}$$

has mean $\mathrm{E}[x] = \bar{x}$ and variance $\mathrm{E}[(x-\bar{x})^2] = \sigma^2$, and so these constants in the gaussian density function have been aptly named.

1.2 Sketch and label your own diagram of the electromagnetic spectrum on a log scale. Label your diagram with the three systems of units: frequency; wavelength; and frequency. Comment on which unit might be the more appropriate to use in each region of the spectrum.

1.3 Explain how the Bayes formula follows from the definitions of marginals and conditionals.

1.4 Prove that a waveform of the form

$$\widetilde{s}(x,y,z,t) = A(x,y)\cos\left(2\pi f_0(t-z/c)+\theta\right)$$

does not satisfy the wave equation

$$\frac{\partial^2\widetilde{s}}{\partial x^2} + \frac{\partial^2\widetilde{s}}{\partial y^2} + \frac{\partial^2\widetilde{s}}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2\widetilde{s}}{\partial t^2}$$

unless $A(x,y)$ is a constant, A, and so is independent of x and y. Conclude that a spatially modulated plane wave satisfying the wave equation does not exist.

1.5 A scalar-valued plane wave has the complex baseband representation

$$s(x,y,z) = A\mathrm{e}^{j2\pi f_0(\alpha x+\beta y+\gamma z)/c+\theta},$$

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press

where A is a complex constant and the direction cosines α, β, and γ are constants. Show that, in general, the sum of two scalar-valued plane waves is not a plane wave. When is the sum a plane wave?

1.6 Answer the following:

(a) Is convolution of finite-energy functions commutative? That is, is $f(x) * g(x)$ equal to $g(x) * f(x)$?

(b) Is convolution of finite-energy functions associative? That is, is $\big(f(x) * g(x)\big) * h(x)$ equal to $f(x) * \big(g(x) * h(x)\big)$?

1.7 The general form of a bivariate gaussian density function on the vector random variable $\boldsymbol{x} = (x, y)$ is

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi \boldsymbol{\Sigma})}} \mathrm{e}^{-(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\dagger} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\bar{\boldsymbol{x}})/2},$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

and $|\rho| \leq 1$. Find the marginals $p(x)$ and $p(y)$ and find the conditionals $p(x|y)$ and $p(y|x)$.

1.8 The general form of a multivariate gaussian density function on the vector random variable $\boldsymbol{x}$ is

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi \boldsymbol{\Sigma})}} \mathrm{e}^{-(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\dagger} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\bar{\boldsymbol{x}})/2},$$

where $\boldsymbol{x}$ is a random vector of length n.

Compute $\mathrm{E}[\boldsymbol{x}]$ and $\mathrm{E}\big[(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])^2\big]$, showing that these are equal to $\bar{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}$, respectively. Can we conclude that these two quantities of $p(\boldsymbol{x})$ are well-named and well-designated as the mean and the covariance matrix?

https://doi.org/10.1017/9781009356190.002 Published online by Cambridge University Press