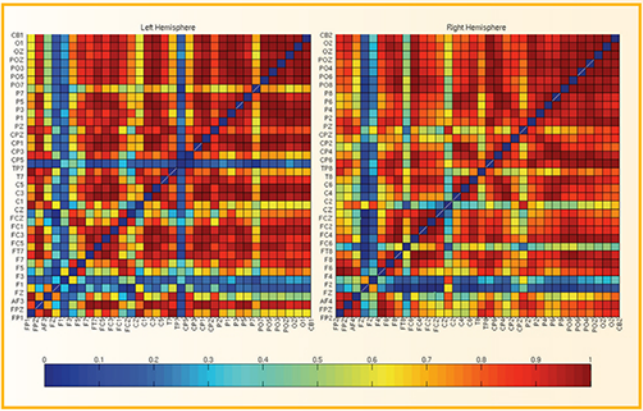
SERIES IN MEDICAL PHYSICS AND BIOMEDICAL ENGINEERING

A BRIEF SURVEY OF QUANTITATIVE EEG



Kaushik Majumdar

CRC Press
Taylor & Francis Group

A Brief Survey of Quantitative EEG

Series in Medical Physics and Biomedical Engineering

Series Editors: John G. Webster, E. Russell Ritenour, Slavik Tabakov, and Kwan-Hoong Ng

Recent books in the series:

A Brief Survey of Quantitative EEG
Kaushik Majumdar

Handbook of X-ray Imaging: Physics and Technology
Paolo Russo (Ed)

Graphics Processing Unit-Based High Performance Computing in Radiation Therapy
Xun Jia and Steve B. Jiang (Eds)

Targeted Muscle Reinnervation: A Neural Interface for Artificial Limbs
Todd A. Kuiken, Aimee E. Schultz Feuser, and Ann K. Barlow (Eds)

Emerging Technologies in Brachytherapy
William Y. Song, Kari Tanderup, and Bradley Pieters (Eds)

Environmental Radioactivity and Emergency Preparedness
Mats Isaksson and Christopher L. Rääf

The Practice of Internal Dosimetry in Nuclear Medicine
Michael G. Stabin

Radiation Protection in Medical Imaging and Radiation Oncology
Richard J. Vetter and Magdalena S. Stoeva (Eds)

Statistical Computing in Nuclear Imaging
Arkadiusz Sitek

The Physiological Measurement Handbook
John G. Webster (Ed)

Radiosensitizers and Radiochemotherapy in the Treatment of Cancer
Shirley Lehnert

Diagnostic Endoscopy
Haishan Zeng (Ed)

Medical Equipment Management
Keith Willson, Keith Ison, and Slavik Tabakov

Quantifying Morphology and Physiology of the Human Body Using MRI
L. Tugan Muftuler (Ed)

A Brief Survey of Quantitative EEG

Kaushik Majumdar

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-1-4398-9616-7 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Majumdar, Kaushik, author.
Title: A brief survey of quantitative EEG / Kaushik Majumdar.
Description: Boca Raton : Taylor & Francis, 2018. | Includes bibliographical references and index.
Identifiers: LCCN 2017017775 | ISBN 9781439896167 (hardback : alk. paper) | ISBN 9781315117256 (ebook)
Subjects: | MESH: Electroencephalography--methods | Computing Methodologies
Classification: LCC RC386.6.E43 | NLM WL 150 | DDC 616.8/047547--dc23
LC record available at https://lccn.loc.gov/2017017775

Visit the Taylor & Francis Website at
http://www.taylorandfrancis.com

and the CRC Press Website at
http://www.crcpress.com

To my parents, who guided me the best, because they guided me the least.

Contents

Preface

Human electroencephalography (EEG) has come a long way since it was first studied by Hans Berger in the 1920s. In the decades that followed, it used to be analyzed by visual inspection, which we can call qualitative studies of EEG. A paradigm shift in the analysis was introduced in the late 1970s, when EEG was started to be analyzed by electronic computers. The subject *quantitative EEG*, or *qEEG*, was born. Digital signal processing (DSP) techniques have been in extensive use in EEG acquisition, storage, and analysis. One big advantage of DSP over analog signal processing (by which EEG used to be recorded and processed in many places until the mid-1990s) is that the digitized signals can be stored in a computer and processed with the help of suitable computer algorithms, whereas pure analog signals can only be stored in a graphical form on a tape and need a specific type of hardware (electronic circuit) for each processing task other than visual means.

In a large majority of cases, EEG researchers follow standard techniques or their variations for digital EEG processing by electronic computers often without going into the technical details of those techniques. The purpose of this book is to elucidate on some of those basic quantitative (i.e., mathematical, statistical, or computational) techniques in as accessible a manner as possible (for this author). Familiarity with twelfth standard mathematics, elementary matrix operations, rudimentary complex numbers, and elementary calculus has been assumed on part of the reader of this text. Familiarity with computer programming, particularly in MATLAB®, although not essential, will be a plus. No background in signal processing is necessary beyond an intuitive idea about what a digital signal is. No knowledge in biology or neuroscience is required.

The target audience of this book includes, but is not limited to, advanced undergraduate or graduate students in computer engineering, cognitive science, psychology, neural engineering, biomedical engineering, electrical or electronics engineering, radiology, neurology, etc. Practicing researchers interested in a deeper understanding of digital EEG processing techniques will find the book quite useful.

The book is divided into nine chapters. Most of the material has been garnered from scholarly articles published in wide-ranging journals from neuroscience, cognitive psychology, physics, signal processing, computer science, applied mathematics, statistics, and others. Some textbooks and monographs, particularly in DSP, cognitive psychology, and EEG turned out to be very helpful. Extensive references have been provided at the end of each

chapter in order to serve the dual purpose of acknowledging the sources, wherefrom materials for this book have been collected, and also to expose the readers to more fundamental and detailed sources for a deeper perusal, if interested. However, the quality of the book being self-contained has not deliberately been compromised.

Chapter 1 deals with the neurobiological origin of scalp EEG and various cortical electrical source models. EEG electrode placement conventions and artifacts have been discussed. Also discussed are data handling software and important frequency bands.

Chapter 2 is devoted to EEG preprocessing techniques. Low-pass and high-pass filtering techniques, principal component analysis (PCA), and independent component analysis (ICA) have been discussed elaborately with a comparative discussion on which ones are suitable for what types of tasks.

Chapter 3 is on the cortical source localization of scalp EEG. The forward problem with various head models has been elaborately described. It is followed by detailed descriptions of various inverse problem-solving algorithms.

Event-related potentials (ERPs) are discussed in the Chapter 4. Various measurement techniques for an ERP are discussed. Then some of the statistical tests to ascertain the effects of ERP over location and time are described.

Chapter 5 is about the binding problem. Synchronization among the EEG signals from various parts of the brain is the most fundamental measurable quantity to study the binding problem by analytical and computational means. Different kinds of phase synchronization measures have been discussed in detail in this chapter. Some other signal similarity measures used in the binding problem have also been deliberated. Lastly, the tricky issue of how two signal similarity measures are to be extended to more than two signals has been discussed.

Chapter 6 is about epileptic EEG. Techniques involving seizure detection and seizure prediction are described. A set of novel heuristics for avoidance of false positives are described. Seizure focus lateralization in scalp EEG is discussed. Processing of interictal (in between two successive seizures) EEG signals mainly for the identification of interictal spikes has been discussed.

The brain–computer interface (BCI) is the topic of Chapter 7. BCI algorithms are mainly about EEG classification into different task categories. First, some special preprocessing techniques for EEG to be used in BCI are described. Then time domain and frequency domain features are discussed. Finally, some machine learning techniques are described for classifying EEG according to BCI tasks.

Chapter 8 gives an overview of fMRI signal acquisition and processing. Magnetic resonance imaging has been discussed, followed by functional

magnetic resonance imaging. The relationship between blood oxygen level-dependent (BOLD) effect and fMRI is discussed. The general linear model for the interpretation of fMRI is explained.

Finally, Chapter 9 deals with the simultaneous acquisition of EEG and fMRI signals. EEG is prone to artifacts in the presence of a high and variable magnetic field. The MR gradient or imaging artifact and ballistocardiogram or heart-rate artifact are the two major artifacts that contaminate the EEG acquisitioned in an MR environment. Both of them have been discussed in detail. Acquisition techniques for minimization of artifacts are described. Interpretation of simultaneous EEG and fMRI is discussed.

Time-frequency analyses of the EEG signals are the most widely used techniques and extensively covered in most existing texts. Here, we have elaborated basic notions of Fourier transforms in Appendix A and wavelet transforms in Appendix B.

Acknowledgments

My foremost acknowledgment goes to the students whom I guided or taught. There were engineers, doctors, mathematicians, statisticians, biologists, and physicists among them. During the course of my interactions with this largely heterogeneous student population, I realized that tremendous interest had been generated to know about the brain, cutting across the disciplines. Most of the material for this book were gathered during my teaching of brain electrical signal processing at the Indian Statistical Institute and the Indian Institute of Science as a guest faculty. At different points of time, my students and colleagues read parts of the manuscript, and their comments helped to improve the book. I particularly thank Pradeep D. Prasad for generating many of the figures including the one in the cover of the book. Also, Ganne Chaitanya, Sonal Sharma, and Anagh Pathak generated some more. I am thankful to all of them. I thank Shivendra Tewari, my former postdoctoral fellow, for his help in getting many of the reference materials that I did not have access to.

The persistence and patience of Aastha Sharma of Taylor & Francis Group remained instrumental all along the writing of the book, from signing the contract in 2011, to finally finishing it in 2016 after missing a couple of deadlines. My mom, who is a retired biology teacher, also took an extraordinary interest in the book and its completion in time.

1

Neurophysiology of the Human Scalp EEG

Every living brain generates electrical potential spontaneously without break. This was detected by Richard Caton of England in the brains of rabbits and monkeys and published his findings in 1875, which is known to be the first recorded case of brain electrical field potential detection. In the 1920s, Hans Berger in Germany started systematically studying it, as it was detectable on the human scalp. It was he who coined the term electroencephalogram (EEG) in 1934 for the recording of the brain electrical field potential. Generally, the term EEG refers to the recording of the scalp potential. In this book, we will remain focused on human EEG.

The brain is composed of two types of cells namely, glial cells and nerve cells, or neurons. Neurons are electrically excitable, which means that in response to specific inputs each neuron can generate electric potential, called action potential. The human brain contains neurons on the order of 10^{11}, about 85% of which are excitatory pyramidal neurons. These are the cells predominantly responsible for generating the EEG. To be more precise, EEG of the human scalp is an indirect effect of the action potential of the tens of thousands of pyramidal neurons firing simultaneously.

Pyramidal neurons are also long projection neurons, which means they can receive input from one region of the brain and send the output through their long axon to another region up to a few centimeters away from the input zone. The remaining, about 15%, neurons of the brain are mostly inhibitory interneurons (there are excitatory interneurons too). Interneurons are those whose input and output remain confined within a small neighborhood. Pyramidal neurons have only a few varieties, whose differences are mostly regarding their shapes, whereas interneurons have much more versatile variations, and one of their major tasks is to keep in check the firing of action potentials by the pyramidal neurons. When this balance is disturbed, neurological disorders like epileptic seizures may occur. Balanced firing activities at the neuronal network levels in the brain as a combined action of excitatory neurons and inhibitors are key to the normal functioning of the brain. We may consider pyramidal neurons to be electric powerhouses in the brain, which become activated by inputs from the sensory world as well as from within the brain. Interneurons constitute sophisticated electronic control switches, which give rise to a

vast number of different output electrical patterns in the generated electric field. These patterns are the representations of the sensory processing and the internal thought processes in the brain. EEG is supposed to capture, at least partly, some of these patterns that are strong enough to be detected on the scalp.

Here are two major caveats: (1) Our brain (or the cortex to be more particular) has six layers, electrical activities in the topmost layer (layer 1) are likely to be captured only in the scalp EEG. Interestingly, layer I does not contain any pyramidal cells. It contains mostly interconnections between the cells. Major inputs come from the pyramidal neurons and branch out to other neurons, including a large number of pyramidal neurons. So, the EEG is generated in the junctions (known as synapses) in between two nerve cells. (2) At any given time our brain is engaged in multiple tasks. So patterns representative of different tasks being executed simultaneously will superimpose, from which separating out a particular pattern very precisely is impossible using the current technologies. Nevertheless, some dominant patterns can be detected with a good degree of certainty, and here exactly lies the usefulness of the EEG. In order to make meaningful observations, computational processing of the EEG signals almost always has to be done under certain neurophysiological constraints. Therefore, a good biological knowledge of the EEG is a must for anyone in the EEG signal-processing community. The current chapter is devoted to this purpose.

1.1 Neural Basis of EEG

The human brain is shown in Figure 1.1. The outermost wrinkled layer is called the cortex or the neocortex. This is what is precisely known as the gray matter, responsible for our intelligence. Only mammals have the neocortex. Not all mammals have a folded neocortex like humans. For example, the mouse neocortex is quite flat. In the case of the mammals, the terms neocortex, cortex, and brain are often used interchangeably. We will follow the same convention in this book. If we imagine a perpendicular axis piercing the brain in Figure 1.1 from top to bottom, the horizontal slices of the brain perpendicular to that axis will be the axial slices. One such slice is shown in Figure 1.2. The thickness of the human neocortex varies from 2 to 4 mm and is divided into six layers as shown in Figure 1.3. The pyramidal neurons responsible for generating the EEG are present in layers III and V. The axons of the pyramidal neurons, through which the action potential propagates, are always perpendicular to the cortical surface (Figure 1.3). Due to the folding of the cortical surface the orientation of this action potential propagation changes with respect to the scalp surface. This creates varied

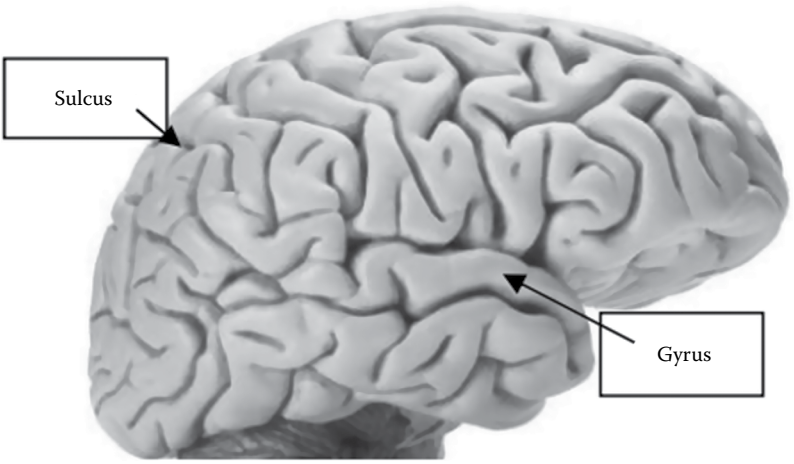


FIGURE 1.1
Human brain. The frontal part is on the left. The outermost wrinkled surface is the cortex. The grooves are known as sulci (sulcus in the singular) and the ridges as gyri (gyrus in the singular).

effects of the electrical activity of a group of locally clustered neurons on the EEG. Also, it matters a lot whether the neuron cluster is located deep inside a sulcus or on top of a gyrus. As can be seen in Figure 1.3, within a small neighborhood, the axons of the pyramidal neurons are parallel to each other and normal to the cortical surface. However, the axons of the interneurons are neither parallel, even within a small neighborhood, nor normal to the cortical surface. This is the reason why interneurons do not make much contribution in the generation of the EEG. However, modeling studies show that stellate cells of layer IV and spiny cells of layer III (both are excitatory interneurons) make a small contribution (Murakami and Okada, 2006).

The cortex is where the EEG is generated, but its rhythm is modulated by the thalamus. Each part of the cortex is connected to the thalamus by feedforward and feedback loops. This enables the thalamus to modulate the frequency of the EEG in each area of the cortex. Oscillations of the electrical signals in the brain also take place due to the interactions between excitatory and inhibitory neurons. Figure 1.4 explains the occurrence of polarity in the EEG. The polarity of the surface EEG depends on the location of the synaptic activity within the cortex.

A thalamic cortical excitatory input in layer IV (left of Figure 1.4) causes a downward deflection at the surface EEG electrode because the EEG electrode is nearer to the source. In contrast, the excitatory input from the contralateral hemisphere in layer II (right of Figure 1.4) causes an upward deflection because it is nearer to the sink (Kandel et al., 2000, p. 915). This is how we can

FIGURE 1.2
Axial slice of the human brain. The gray matter or the cortex is the part where the sophisticated brain activities take place. It is also the source of the EEG. White matter consists of myelinated axons of the long projection neurons. They are responsible for communication between different cortical areas.

describe the contribution of synaptic potentials to the dendritic tree of single pyramidal neurons to the EEG collected from the scalp electrodes.

1.1.1 Dipole Source Model

For the purpose of quantitative analysis, the cortical sources of scalp EEG are mathematically or computationally modeled in either of two different ways, namely, *the dipole source model*, or *the equivalent current dipole model*, and *the distributed source model*. In this section, we discuss the dipole source model. In the next section we will take up the distributed source model.

In Figure 1.4 it has been shown how a pyramidal neuron oriented vertically to the cortical surface can alter the polarity depending on whether the input is received at the proximal or apical dendrites. In the case of multiple inputs, summation of all apical inputs and summation of all proximal inputs are to be considered. If in a small region of the cortex there are several pyramidal neurons with the same orientation, then that region will have

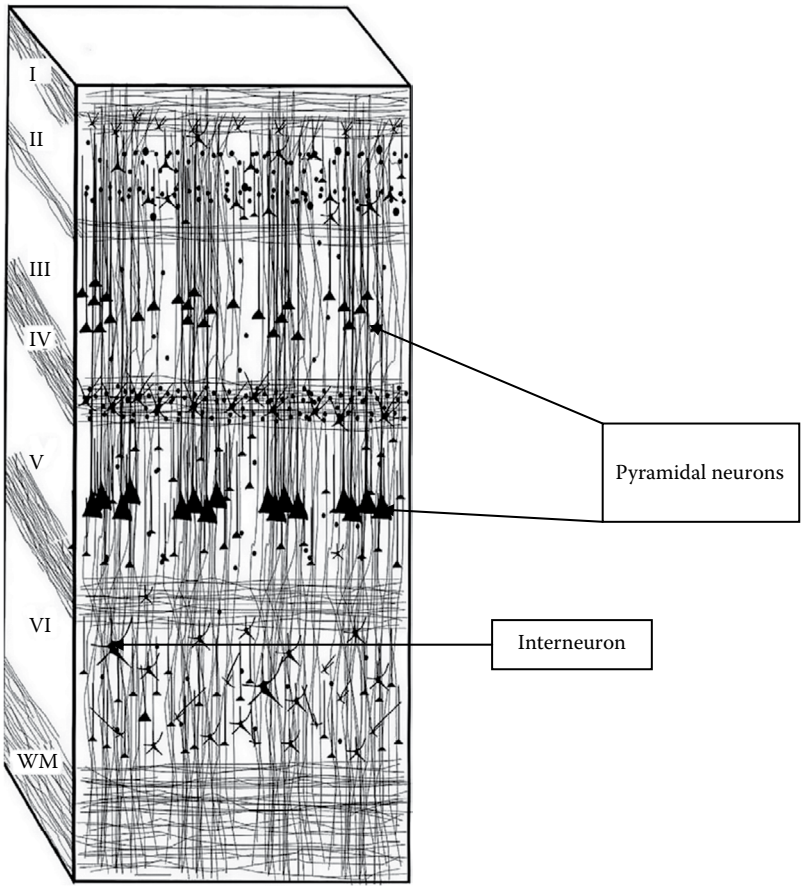


FIGURE 1.3
Cortical layers I–VI: WM = white matter. Note that the axons of the pyramidal neurons are oriented vertically to the cortical surface, ensembles of which within a tiny columnar region constitute a dipole. (Adapted from Bentivoglio, M. et al., *Epileptic Disord.*, 5, S27, 2003.)

a strong unidirectional polarity. The human cortex has a packing density of 10^4 neurons/mm^3, 85% of which are pyramidal neurons. This means there are 8500 pyramidal neurons in 1 mm^3 of cortical volume. Since the human cortex is never more than 4 mm thick and pyramidal neurons (their soma to be more precise) are confined within the layers III and V, we can reasonably assume that there are on average 8500 pyramidal neurons in each mm^2 of the cortical surface, which are vertically oriented to the surface as shown in Figure 1.3 (Fischl and Dale, 2000). Even if 1% of them are oriented in the same direction without too many neurons in the same region with the opposite orientation, then there will be 85 neurons with the same orientation within 1 mm^2, which is a good density for a significant dipolar current source. Obviously, here the dipole consists of 85 pyramidal neurons with the same orientation. In fact, 108

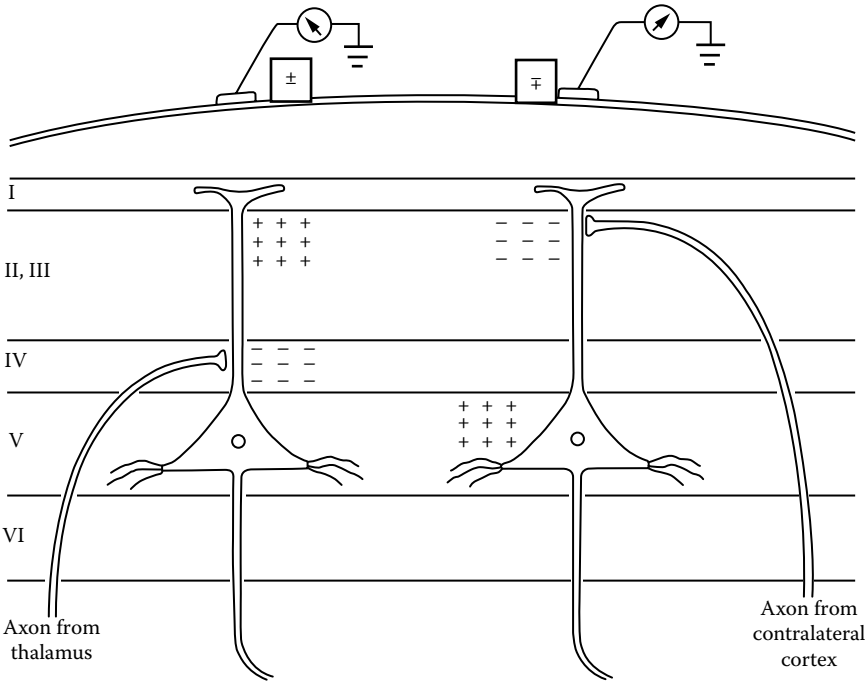


FIGURE 1.4
Layer V pyramidal neurons (triangle with a small circle inside as the soma) take input from all six layers of the cortex. Axons from neurons of the thalamus and the cortex, depending on their projection on the apical (far from soma) or proximal (close to soma) dendrites of the pyramidal neuron, depolarize (negative sign) or hyperpolarize (positive sign) the synapses at the axon end feet. This in turn determines the polarity of the scalp potential. (Modified from Olejniczak, P., *J. Clin. Neurophysiol.*, 23(3), 186, 2006.)

neurons in a cortical area of 6 cm^2 can create a visible EEG effect on the scalp (Olejniczak, 2006). The orientation of this dipole with respect to the scalp surface will depend on the fold of the cortex in which it is situated (Figure 1.5).

Apart from the orientation, the depth of its location (whether deep inside a sulcus or on the top of a gyrus) will also affect its contributing strength in the EEG (Figure 1.6). We will discuss about quantitative modeling of dipoles and their effects on the EEG in Chapter 3.

1.1.2 Distributed Source Model

The dipole source model, although quite appealing, has one important limitation. Our brain is never idle and almost all the parts are active for almost all of the time. So every part of the brain is supposedly contributing to the EEG at any point of time. In this sense the number of sources is very large. In fact

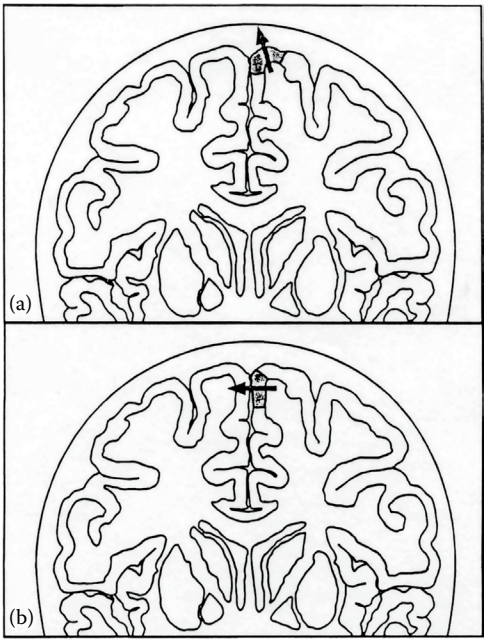


FIGURE 1.5
Due to the different orientations the dipole source in panel (a) will have a greater effect on the scalp electrodes than the dipole source in panel (b).

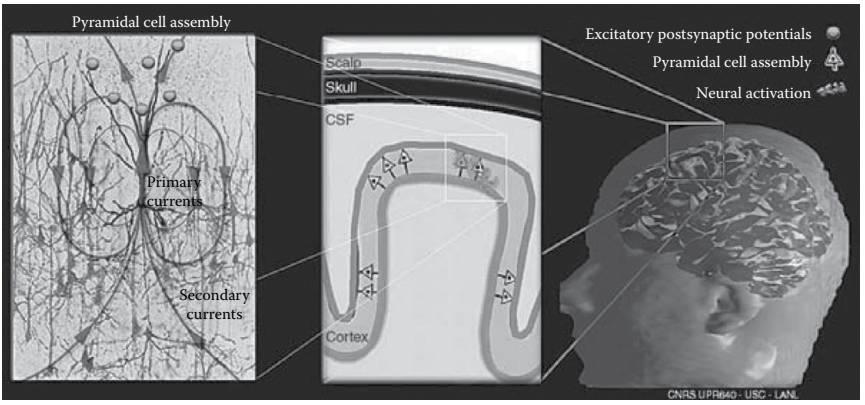


FIGURE 1.6
The equivalent current dipoles on the cortical surface (middle panel). Their generation (left panel) and a patch of activated cortical surface inside the head (right panel) have also been shown. (Taken from Bailet, S. et al., *IEEE Signal Proc. Mag.*, 18(6), 2001.)

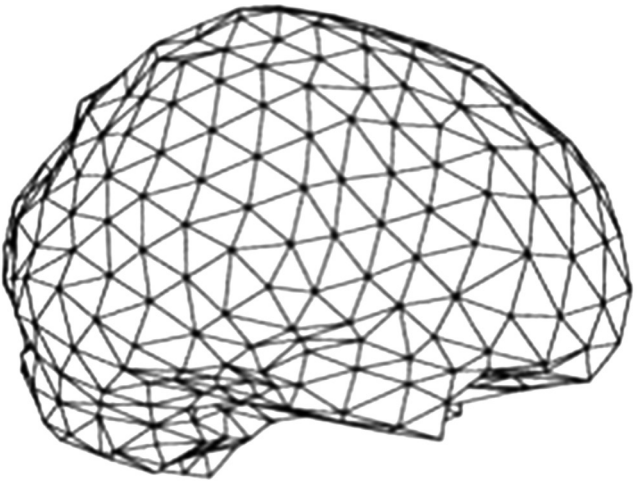


FIGURE 1.7
Triangular wire mesh diagram of the human cortical surface. In the distributed source model, each vertex point of a triangle is a source of electric field contributing to the EEG. (Adapted from Hallez, H. et al., *J. Neuroeng. Rehabil.*, 4, 46, 2007, available at: http://www.jneuroengrehab.com/content/4/1/46.)

they are distributed all over the cortical surface. In computational modeling cortical surfaces are modeled as a geometric triangular mesh (Figure 1.7). Each vertex of the triangle is a source of an electric field due to the excitatory postsynaptic potential (EPSP) (in the dendritic arbors in layers I and II) contributing to the EEG signal at each scalp channel, when sufficiently strong. The number of points in the cortical triangular mesh diagram may range from several thousands to several tens of thousands. Dipole modeling of such a large number of sources becomes extremely difficult. The model is simplified by compromising with the orientation. In a distributed source model, sources are assumed to be distributed all over the cortical surface and are oriented normal to the surface. We will return to the distributed source model again in Chapter 3.

1.2 Tissue Impedance

Biological tissues have a high resistivity, which is often different at different parts of the same tissue. How the electric field in the cortex gives rise to scalp potential in the EEG channels depends a lot on the tissues lying

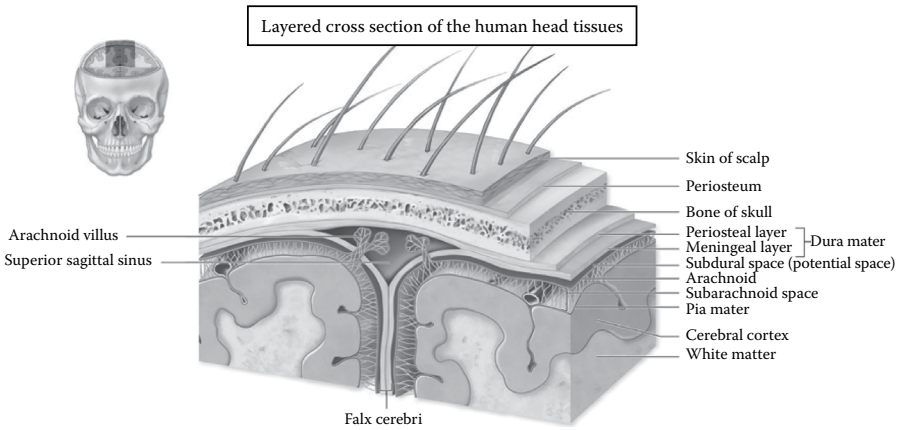


FIGURE 1.8
There are at least nine clearly identifiable tissue layers between cortical electric sources and scalp electrodes; the impedance of each is likely to affect the EEG. (Courtesy of Kandel, E.R. et al., *Principles of Neural Science*, 4th ed., McGraw Hill, New York, 2000.)

between cortical and scalp surfaces. Nine different tissue layers between cortical and scalp surfaces have been shown in Figure 1.8. There is also cerebrospinal fluid (CSF) affecting the generation of EEG. For computational modeling usually three major layers are considered namely, cortex, skull, and scalp. Since the sources sit inside the cortex its impedance matters. The bony tissue of the skull by far has the largest impedance of all. Also, its impedance varies significantly across the surface, because of its nonuniform thickness and porousness. The tissue impedance varies from individual to individual and even in the same individual with age. The tricky issue of taking care of tissue impedance in computational modeling will be discussed in Chapter 3.

Impedance of the head tissues is usually measured postmortem and widely different values have been reported in the literature. Tissue impedance is supposed to vary from a living body to a dead body. It is possible to measure the impedance of the head tissues of a living subject using a method known as *electrical impedance tomography* (EIT). It can be done by measuring currents and voltages on the tissue boundary due to small applied electrical stimulations, which are completely harmless, and determining tissue impedance at different locations as an inverse problem (Borcea, 2002). Usually four electrodes are used to measure the impedance. Two of the electrodes are for applying the stimulation current and two for measuring the consequent voltage difference. Nowadays commercial toolkits for this purpose are also available.

1.3 Artifacts

Although the EEG is designed to record cerebral activity, it also records electrical activities arising from sites other than the brain. The recorded activity that is not of cerebral origin is termed as an *artifact* and can be divided into physiologic and extraphysiologic artifacts. While physiologic artifacts are generated from the patient, they arise from sources other than the brain (e.g., body). Extraphysiologic artifacts arise from outside the body (e.g., equipment, environment) (Benbadis and Rielo, 2008).

1.3.1 Physiologic Artifacts

Muscle artifacts: The artifacts generated due to muscle contractions (particularly the jaw muscles) are the most common among all the artifacts in the EEG. Usually muscle artifacts have a high frequency (50–100 Hz) and they are of shorter duration compared to brain activities. They can be differentiated from cerebral activities by frequency, duration, and morphology (Figure 1.9). These artifacts are also known as electromyograms or EMG artifacts.

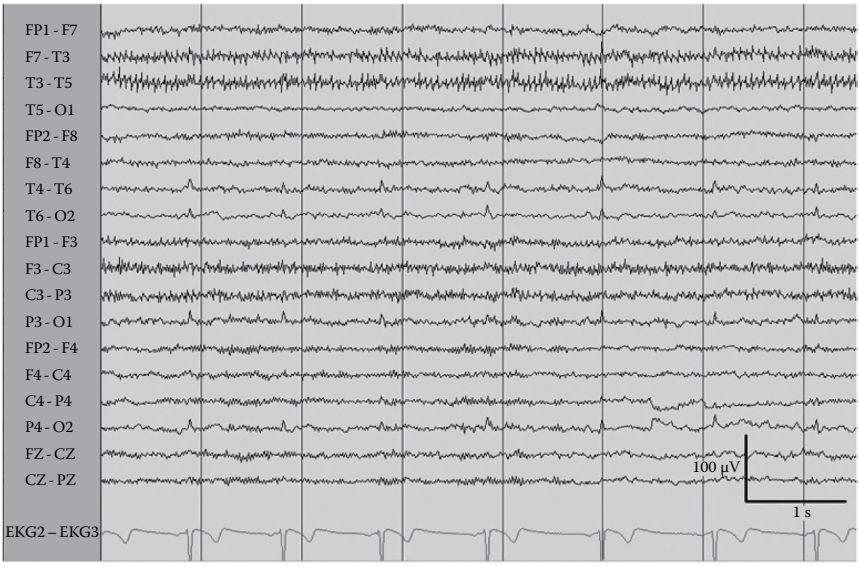


FIGURE 1.9
Muscle artifacts with high frequency and amplitude are clearly visible in the F7–T3 and T3–T5 electrodes. The bottommost electrode records the EKG and the EKG artifacts are clearly identified to match with the rhythmicity of the heart rate. (Taken from Benbadis, S.R. and Rielo, D., EEG artifacts, *eMedicine*, 2008, available at http://emedicine.medscape.com/article/1140247-overview.)

Tongue and eyeball artifacts: Tongue and eyeballs are the most mobile organs within the head and each acts as an electrical dipole (the tip of tongue is negatively charged relative to the base and the cornea of eye is positively charged relative to the retina). Their movements, therefore, produce a detectable potential in the scalp. However the frequency remains low. Tongue-movement artifacts are characterized by broad potential fields that decrease from the frontal to the occipital areas. The frequency is variable but usually remains below 5 Hz. The eyeball movement artifacts also diminish from the frontal to the occipital areas and are steeper than the tongue-movement artifacts. Eye-movement artifacts are more important to deal with than tongue-movement artifacts because of the importance of EEG in sleep studies (eye-movement artifacts will be substantial during the rapid eye movement [REM] stage of sleep). When the eyeball rotates, the dipole orientation changes and potential varies creating an artifact with a frequency ranging from 4 to 20 Hz (Table 1 in Keren et al., 2010). This is detectable by any electrode near the eye, particularly F7 and F8 (see Section 1.4 for information on electrode placement).

Eye-blink artifact: An eye blink can last from 200 to 400 ms and can produce an EEG artifact of magnitude 10 times that of cortical signals (Joyce et al., 2004) (Figure 1.10). The majority of these artifacts propagate through the superficial layer of the face and the head, which decrease rapidly with the distance from the eyes. The frequency of the blink artifacts usually ranges from 0 to 4.5 Hz (Table 1 in Keren et al., 2010).

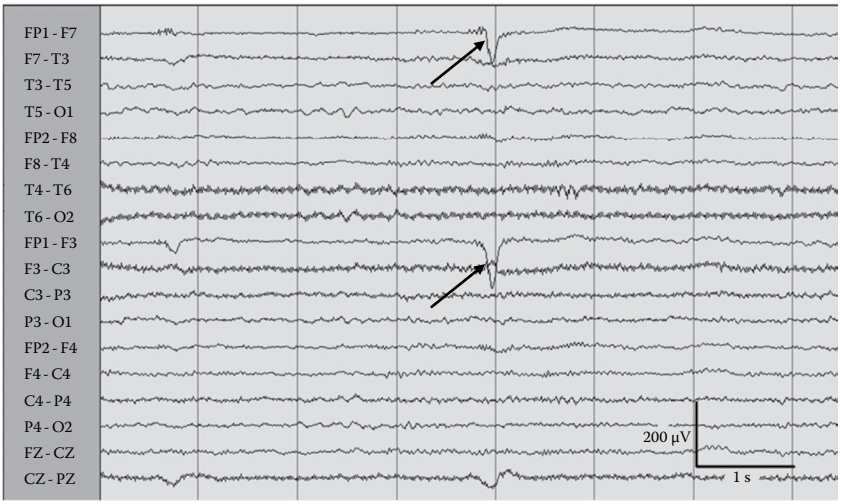


FIGURE 1.10
Eye-blink artifacts indicated by arrows in the EEG of a subject whose one eye is made of glass. Therefore only F7 corresponds to the blink but F8 does not. (Adapted from Benbadis, S.R. and Rielo, D., EEG artifacts, *eMedicine*, 2008, available at http://emedicine.medscape.com/article/1140247-overview.)

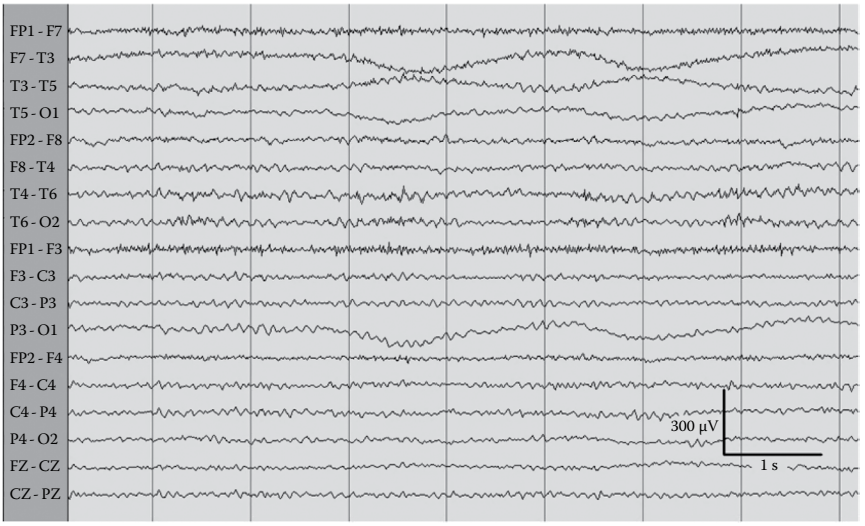


FIGURE 1.11
Slow-wave sweat artifacts are visible in channels F7–T3, T3–T5, T5–O1, and P3–O1. (Taken from Benbadis, S.R. and Rielo, D., EEG artifacts, *eMedicine*, 2008, available at http://emedicine.medscape.com/article/1140247-overview.)

EKG artifact: Also known as the ECG (electrocardiograph) artifact, this is the artifact produced due to the rhythmic motion of the heart. It affects the EEG more if the subject has a wider and shorter neck. These artifacts are easily identified by their rhythmicity and coincidence with the heart rate or ECG tracer (Figure 1.9).

Pulse artifact: If an EEG electrode is placed over a major blood vessel, then a slow-wave artifact can be detected in that electrode whose frequency will be equal to the heart rate.

Sweating artifact: Sweat changes skin impedance and thereby affects the electric field in the skin–electrode junction leading to changes in the electric potential at the electrode. It usually produces a very slow-frequency artifact of about 0.5 Hz (Figure 1.11).

1.3.2 Extraphysiologic Artifacts

Electrode artifacts: The electrode contact points are one of the common sources of artifacts. They remain confined to the specific electrode(s) and appear as sharp waveforms due to sudden impedance change. In general, sharp transients that occur in single electrodes (not field) should be considered as artifacts unless proved otherwise. In some cases the impedance changes may be less abrupt and the artifact may look like a low-voltage arrhythmic wave of frequency less than or equal to 4 Hz.

Line noise: Every EEG acquisition and amplification system is powered by a 220/110 V alternating current. The frequency of this a.c. supply is either 50 (such as in India) or 60 (such as in the United States) Hz. Proper earthing of the subject's body is one convenient way to avoid this artifact (however, for reasons of safety a human body is never directly connected to earth during EEG acquisition). Since the frequency remains very steady, it can be easily removed by filtering (see Chapter 2 for more detail).

Apart from the artifacts listed above there might be artifacts due to the movement of persons close to the subject at the time of recording of the EEG. Also, running of an electric motor (such as fan and air conditioner) nearby may send electrical signals strong enough to contaminate the EEG. Touching or hitting EEG electrodes can produce odd waveforms. In Chapter 2 we will deal with the removal of artifacts from the EEG.

1.4 Electrode Placement Systems

In as early as the first International EEG Congress held in London in 1947, it was recognized that a standard method of placement of electrodes used in EEG was needed. H. H. Jasper introduced the first electrode placement system, known as the 10–20 electrode system, in 1958. Since then the 10–20 system has become the de facto standard for the clinical EEG. It has also proven to be very useful in the study of event-related potential (ERP) in the nonclinical settings. However, the advancement of topographic methods to study spontaneous and evoked potentials, and the advancement of multichannel EEG hardware systems necessitated the standardization of a larger number of channels. Therefore, in 1985, an extension of the original 10–20 system was proposed, which involved an increase of the number of electrodes from 21 (in the 10–20 system) to 74. This extension of the 10–20 system of electrode placement has come to be known as the 10–10 system or 10% system. It has been well accepted and currently recognized as the standard by the American Electroencephalographic Society and the International Federation of Societies for Electroencephalography and Clinical Neurophysiology (Oostenveld and Praamstra, 2001). More recently, a placement system for up to 345 channels has been proposed (Oostenveld and Praamstra, 2001). In this book, however, we will describe only the conventional 10–20 and 10–10 systems.

1.4.1 10–20 System

The standard placement recommended by the American EEG Society for use in the international 10–20 system is for 21 electrodes, but the system is designed to allow the use of additional electrodes with predictable and easily repeatable

placement. The standard numbering system in the 10–20 system places odd-numbered electrodes on the left and even-numbered electrodes on the right, with the latter designating the anatomic area (Figure 1.12). The 10 and 20 refer to the fact that the actual distances between adjacent electrodes are either 10% or 20% of the total front–back or right–left distance of the skull. The letters identify different lobes: F = frontal, P = parietal, C = central, O = occipital, z = zero (midline). Odd indices appear on the left and the even ones on the right.

Electrode positions are measured with respect to two anatomical landmarks namely nasion (between forehead and nose) and inion (lowest point of the skull from the back of the head) (Figure 1.12). The distance between nasion and inion is measured along the midline (let us assume it is 40 cm). Along this line, the frontopolar point Fpz, is marked at 10% above the nasion (i.e., 4 cm above the nasion, it will look below the nasion in Figure 1.12). Frontal (Fz), central (Cz), parietal (Pz), and occipital (Oz) points are marked at intervals of 20% of the entire distance (i.e., 8, 16, 24, and 32 cm, respectively, below Fpz) leaving 10% for the interval between Oz and inion. The midline points Fpz and Oz routinely do not receive any electrode.

Along the horizontal line joining T3 and T4, the position for the central points C3 and C4 and the temporal points T3 and T4 are marked 20% and 40%, respectively, from the midline. The circumference of the head is measured from the occipital point (Oz) through the temporal points T3 and T4 and the frontopolar point (Fpz). The longitudinal measurement for Fp1 is located on that circumference, 5% of the total length of the circumference to the left of Fpz. The longitudinal measurements for F7, T3, T5, O1, O2, T6, T4,

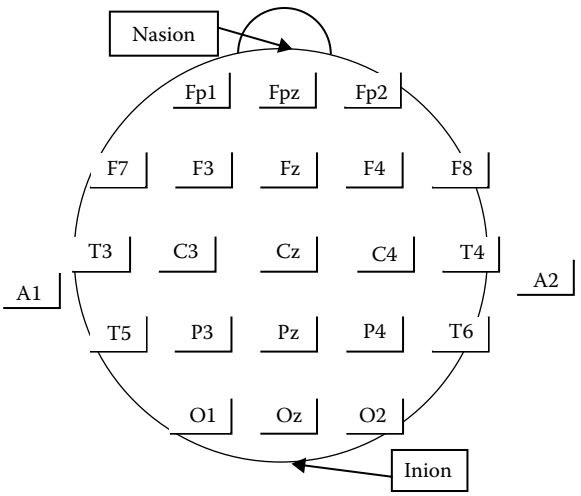


FIGURE 1.12
International 10–20 electrode placement system. Nasion and inion are two anatomical landmarks with respect to which electrode positions are measured. A1, A2 are referential electrodes on the ear lobes. Fpz and Oz conventionally do not receive any electrode.

F8, and Fp2 are at the distance of 10% of the circumference. The F3 electrode is placed at the point of intersection of two lines drawn by joining Fp1 to C3 and F7 to Fz. Similarly F4 is placed on the right side. The P3 electrode is placed at the point of intersection of two lines drawn by joining O1 to C3 and T5 to Pz. Similarly the P4 electrode is placed on the right side.

1.4.2 10–10 System

The 10–10 placement of electrodes is based on landmarks on the skull, namely the nasion (Nz), the inion (Iz), and the left and right pre-auricular points (LPA and RPA). If Nz and Iz are joined by an arc passing through the left ear and by another arc passing through the right ear, LPA will fall on the middle of the arc passing through the left ear, between Nz and Iz, which will be very close to the left ear. RPA is located in the same way. The first step is to form the line from Nz to Iz, approximately over the vertex (Cz). To determine the location of the vertex, the contour from LPA to RPA is also passed over the vertex. These two contours should intersect at 50% of their length, and the point thus obtained is the exact vertex Cz.

Along the sagittal Nz–Iz scalp contour over the vertex, the positions Fpz, AFz, Fz, FCz, Cz, CPz, Pz, POz, and Oz are marked at 10% distances along this anterior–posterior contour (Figure 1.13). With the position of Cz at 50% along this contour, corresponding to the vertex, the position of Oz is at a distance of 90% from Nz and 10% from Iz.

Along the coronal LPA–RPA scalp contour over the vertex, the positions at 10% above the LPA and the RPA are marked. These positions are necessary to determine the horizontal contours over the left and right temporal lobe.

A horizontal circumferential contour is determined over the left temporal lobe from Fpz to Oz, through the location that was marked at 10% above LPA. Along this contour, the positions Fp1, AF7, F7, FT7, T7, TP7, P7, PO7, and O1 are marked at 10% distances (Figure 1.13, innermost circle). The circumferential contour over the right temporal lobe is determined in the same fashion from Fpz to Oz over the location 10% above RPA, and the positions Fp2, AF8, F8, FT8, T8, TP8, P8, PO8, and O2 are marked at 10% distances.

A coronal contour is determined from left to right from location T7 to T8, through the location Cz. Along this contour at a fraction of 1/8 from the total distance, the positions C5, C3, C1, Cz, C2, C4, and C6 are marked. This is conveniently done by bisecting the contour twice on each hemisphere, dividing it into halves and quarters. Similar coronal contours are determined along FT7-FCz-FT8, T7-Cz-T8, TP7-CPz-TP8, P7-Pz-P8, and positions at fractions of 1/8 along these contours are similarly marked. The labels of these locations are described in Figure 1.13.

A further coronal contour is determined left to right over the anterior part of the head, going from location AF7 to AF8 and through location AFz. The distance along this contour for each hemisphere is bisected, giving locations AF3 and AF4. Following this, the bisection is repeated and the additional positions are marked. This gives along this contour, at fraction of 1/8 of the total distance,

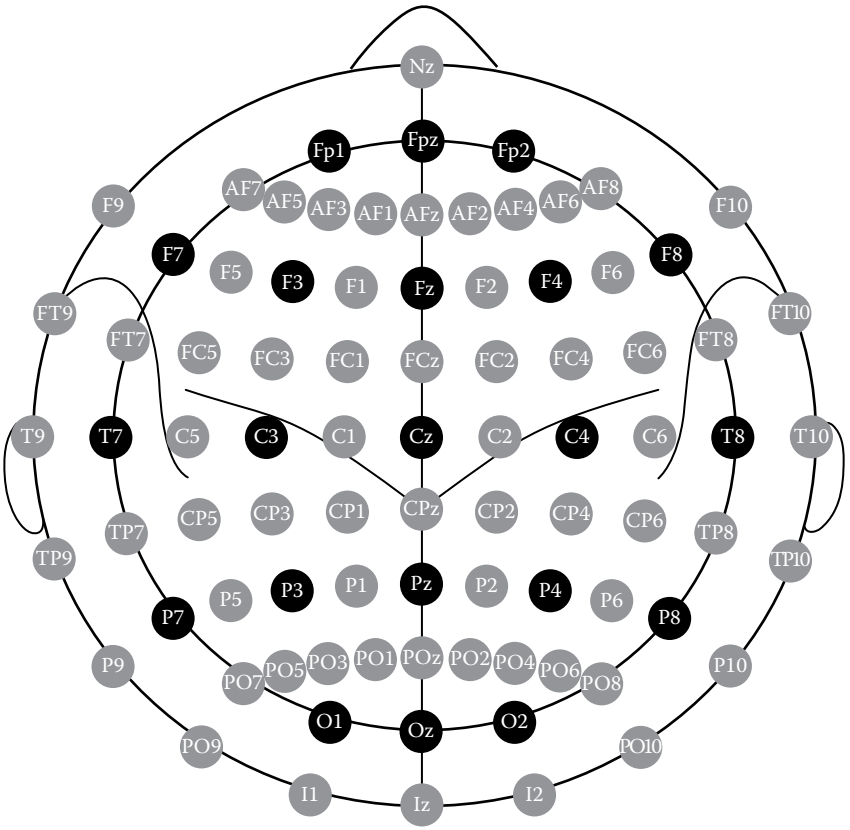


FIGURE 1.13
10–10 electrode placement system. The black circles indicate original 10–20 electrode placement system. The gray circles indicate the extension to the 10–10 system. Nz = nasion and Iz = inion. (Taken from Oostenveld, R. and Praamstra, P., *Clin. Neurophysiol.*, 112, 713, 2001.)

the locations AF5, AF3, AF1, AFz, AF2, AF4, and AF6. The same procedure is performed for the coronal contour from PO7 to PO8 through POz, giving the locations with labels as shown in Figure 1.13. Although the positions AF1, AF2, AF5, AF6, PO1, PO5, PO2, and PO6 are not explicitly mentioned and displayed in the standard 10–10 system, they are already present implicitly.

1.5 Noise

In EEG analysis the words noise and artifact are used interchangeably. However, let us make a small distinction over here. On the one hand, by artifact in this book we will understand the signals of noncerebral nature

present in the EEG, whose sources we are sure of. On the other hand, by noise we signify the noncerebral signals present in the EEG whose sources are not known clearly. Noise is often treated as random and therefore has positive and negative amplitudes almost in equal numbers for a sufficiently long duration. In many analyses, therefore, the EEG signals across trials (see the next section for the meaning of 'trial') are averaged in the hope that the positive and negative parts of the noise will cancel each other, and in the trial-averaged signal the signal-to-noise ratio (SNR) will be higher than that in the single trials. However this is not true in general. Note that SNR can be defined in many different ways. One simple way to define it is

$$SNR = \frac{A_{signal}^2}{A_{noise}^2}, \tag{1.1}$$

where

A_{signal} is amplitude of the signal (desired signal)
A_{noise} is the amplitude of the noise (the unwanted signal)

The term noise is relative and depends on the application. Someone's noise may be someone else's signal and vice versa. In a slightly more sophisticated way SNR is defined in decibel (dB) units as

$$SNR_{dB} = 10\log_{10}\left(\frac{A_{signal}^2}{A_{noise}^2}\right). \tag{1.2}$$

Equation 1.2 is probably the most widely used definition of SNR.

1.6 Data Representation

Typical EEG recordings may continue from tens of minutes to days on end. The number of channels may be from 16 to 300, with the sample frequency (rate at which a continuous or analog signal is discretized) ranging from 200 to 20,000 Hz (that is, 200–20,000 values of the analog signal are sampled per second). During the recording, the subject or the patient (EEG is most notably recorded in the epileptic patients for clinical diagnosis, monitoring, and evaluation) may be asked to perform certain tasks, such as visual object identification, experiencing mild electrical stimulation at a particular body part, listening to a musical tone, etc. Each instance of a task is called a trial and its duration is usually fixed for all the instances.

Now say a person is receiving mild electrical stimulations on the median nerve of the right palm. He has received a total of 490 pulses at equal

intervals and with instantaneous duration. The EEG has been extracted from the recording from 30 ms before the stimulus to 50 ms after the stimulus for each of the 490 pulses. Say the sample frequency is 5000 Hz, and there are 64 scalp channels to collect the EEG. Then there are $64 \times 490 \times (5 \times 80)$ number of data points. In this case, one single file may be more than 45 MB long in some format. Continuous EEG for days on end may generate data files several hundreds of GB long in any format.

The format of EEG data depends exclusively on the acquisition system. Different vendors of the system store data in different formats. Fortunately, software is available to convert the raw EEG data files into a format convenient for processing by electronic computers. This is indeed a great boon to the quantitative EEG community. EEGLAB is one such open source software that is freely available and regularly updated (EEGLAB, 2017). Although EEGLAB is free it works only on a MATLAB® platform, which is not free. However the compiled version of EEGLAB does not require MATLAB although scripting capabilities are more limited.

EEGLAB is a very user-friendly GUI-based application. It runs on Linux, Unix, Windows, Mac OS X, and supports EEG file formats in binary, MATLAB (.mat) ASCII (.asc and .txt), Neuroscan, EGI, Snapmaster, European Standard BDF, and Biosemi EDF (Delorme and Makeig, 2004).

EEGLAB uses a single structure ('EEG') to store data, acquisition parameters, events, channel locations, and epoch information as an EEGLAB dataset, in the form of a MATLAB cell array, which can be accessed directly from the MATLAB command line. Text files containing event and epoch information can be imported via the EEGLAB menu. The user can also use the menu to import event and epoch information in any of several file formats (Presentation, Neuroscan, ASCII text file), or can read event marker information from the binary EEG data file (as in, for example, EGI, Neuroscan, and Snapmaster data formats). The menu then allows users to review, edit, or transform the event and epoch information. Event information can be used to extract data epochs from continuous EEG data, select epochs from EEG data epochs, or to sort data trials to create event-related potential (ERP) (see Chapter 4) image plots. EEGLAB also provides functions to compute and visualize epoch and event statistics (Delorme and Makeig, 2004).

1.7 Frequency Bands

For convenience of analysis from clinical as well as from the cognitive science point of view EEG signals have been divided into several frequency ranges. They are delta (0–4 Hz), theta (4–8 Hz), alpha and mu (8–13 Hz), beta (13–30 Hz), and gamma (30–80 Hz). Different rhythms or frequencies in the EEG are

modulated by the following three neurophysiological factors (Pfurtscheller and Lopes da Silva, 1999):

1. Intrinsic membrane properties of the neurons and the dynamics of the synaptic processes.
2. The strength and extent of the interconnections between the network elements, most often formed by feedback loops. Different kinds of feedback loops can be distinguished, involving thalamocortical or corticocortical either at short or long distances.
3. The modulating influences from general or local neurotransmitter systems.

1.7.1 Delta

Delta band EEG is known to be a low-frequency, high-amplitude wave, which among others is associated with slow wave sleep (SWS). SWS in turn is associated with declarative memory consolidation (Marshall et al., 2006). Declarative memory is the memory of events in one's life. This indicates that delta wave may have implication in memory consolidation. In the adult brain, delta activity is originated in the frontal cortex, hippocampus, and thalamus. Delta EEG is important in sleep research because it is associated with the deepest stage of sleep known as nonrapid eye movement (NREM) sleep. K-complex is a distinctive pattern of sleep EEG, which precedes delta waves.

Although delta wave is associated with the SWS, it is implicated in the attention paid to the performance of a mental task (Harmony et al., 1996), such as a complicated arithmetic calculation. In fact, signal power at delta frequency band increases with the increasing difficulty of the mental task. In tasks where attention to the external environment needs to be paid, the delta power goes down. It goes up only when attention to internal processing needs to be enhanced (for example, during mental calculation). It is interesting to note that very slow, low-voltage delta waves have been associated with schizophrenia (Itil, 1977).

Delta sleep is mediated by thalamocortical circuits, and delta wave deficits during sleep have been observed in patients with schizophrenia. Negative symptoms have been reported to reflect frontal lobe dysfunction and to be associated with decreased delta wave sleep. In the central region and other brain areas, the same deficiency of schizophrenia patients compared to the normal control has been observed (Sekimoto et al., 2011).

1.7.2 Theta

The theta band (intracranial) EEG in humans may be generated either from the cortex or from the hippocampus, and these two forms may be independent of each other (Cantero et al., 2003). However, recent research on rat brain

has identified high coherence between theta rhythms generated in the hippocampus and the prefrontal cortex during spatial learning (Benchenane et al., 2010). Since the hippocampus is located quite deep inside the brain, the theta rhythm generated by it is unlikely to be detected on the scalp. Although no consensus has yet emerged on the theta band's specific behavioral correlates, they are most consistently present during REM sleep and during various types of locomotor activities described by the subjective terms 'voluntary', 'preparatory', 'orienting', or 'exploratory'. In general, theta waves are absent in immobile animals, but short epochs of theta trains can be elicited by noxious, conditioned stimuli. A different type of theta wave on scalp EEG has been observed during meditation (Banquet, 1973). Also, the theta wave is associated with the mnemonic processes in our brain (Buzsaki, 2002), where the step-by-step detail of a process up to the sequential details has to be taken care of. Clearly theta wave plays a role in the working memory. It has also been implicated in Hebbian strengthening of hippocampal synapses and in both short-term and long-term potentiation (STP and LTP, respectively). The induction of LTP occurs at the positive phase of the theta rhythm. The strength of the induced LTP increases linearly with the theta power. Pharmacological manipulations demonstrate that drugs that decrease theta activity also block learning, whereas drugs that promote theta rhythm (and thus enhance the LTP) also facilitate learning (Klimesch, 1999).

There are interneurons in the hippocampus and prefrontal cortex that intrinsically oscillate in the theta-band frequency (Buzsaki, 2002). During high–theta band coherence in the prefrontal cortex (particularly during learning novel spatial locations), the local field potential (LFP) in both the hippocampus and the prefrontal cortex may be in high synchrony. It has been observed that rodents that have theta rhythms in the hippocampus and the entorhinal cortex acquire high power during learning and memory retrieval.

1.7.2.1 Alpha

It has an amplitude of 20–100 μV, best seen with eyes closed, and attenuates with eye opening. However, those who practice meditation may have high-amplitude alpha rhythms EEG even after opening the eyes following meditation (Banquet, 1973). The origin of alpha wave is posterior dominant with maximum part at the anterior, particularly in the occipital lobe, and it may be more widespread. It has a rhythmic, regular, and waxing and waning morphology (Benbadis, 2006).

Within the alpha frequency range the EEG power is positively related to cognitive performance and brain maturity, whereas the opposite holds true for the theta frequency range. During actual task demands, the extent of alpha power suppression is positively correlated with cognitive performance (and memory performance in particular), whereas again the opposite holds true for the theta band (Klimesch, 1999). Regular alpha waves do not form in humans until 3 years of age (Niedermeyer, 1997). It is the dominant

frequency in the human scalp EEG of adults. It is manifested by a 'peak' in spectral analysis (Klimesch, 1999).

Alpha frequency is positively related to cognitive performances. Large power in the range of the upper alpha band but small power in the theta frequency range indicates good cognitive performance. The power of alpha frequency increases from early childhood to adulthood but then decreases with increasing age or age-related neurological diseases. The amplitude of the alpha frequency is lowered in subjects with dementia, as well as patients in many other types of neurological disorders (Klimesch, 1999). The alpha frequency has a significantly higher amplitude in subjects with good memory performance as compared to age-matched controls with bad memory performance. The amplitude of alpha frequency is positively correlated with the speed of information processing by the subject. Theta power decreases and alpha power increases from early childhood to adulthood. Theta power increases and alpha power decreases in the late part of the lifespan. Theta power is enhanced and alpha power is lowered in subjects with a variety of different neurological disorders as compared to age-matched controls (this holds true not only for demented subjects but also for children with reading/writing or spelling difficulties). During the hypnogenic state (transition from waking to sleeping) when the ability to respond to external stimuli decreases, theta power increases and upper alpha power decreases (Klimesch, 1999).

1.7.2.2 Mu

Mu rhythm (mu stands for motor) has an arch-shaped wave morphology, which looks like an 'm', and appears over the motor area (centroparietal area). It is usually asymmetric and asynchronous between the two hemispheres and may be unilateral. Its amplitude is in the range of alpha rhythm. It becomes selectively enhanced or synchronized, and suppressed or desynchronized during motor-related task demands (Pfurtscheller et al., 2006).

1.7.3 Beta

It is generated mostly in the frontocentral part but the location is variable. The morphology of the band is usually rhythmic, waxing and waning, and symmetric. It is usually in the range of 5–20 μV (Benbadis, 2006).

Beta EEG is associated with our waking state in general, and cognitive and emotional processes in particular (Ray and Cole, 1985), although beta activity is enhanced by 18–25 Hz during sleep stages 1 and 2 and tends to decrease during deeper sleep stages (Benbadis, 2006). Hyper beta activity at the sleep onset has been associated with insomnia (Perlis et al., 2001). Since cortical waves follow the $1/f$ power law, where f is frequency (Buzsaki, 2004), it is reasonable that the signal power associated with the beta EEG band will be quite predominant all over the scalp surface. Due to the difference in thickness

between the two hemispheric skulls, the amplitude of beta wave may be up to 35% different on the two hemispheres of the head (Benbadis, 2006).

Enhanced high-frequency beta (16–25.5 Hz) activities have been observed in schizophrenia patients (Itil, 1977; Michel et al., 1993). Opposite effects were found in the two beta ranges of 20.5–22.5 and 23.0–25.5 Hz (Michel et al., 1993).

1.7.4 Gamma

In earlier days analog EEG used to be recorded on paper tape with an ink pen moving up and down with strokes of (amplified) brain electrical signals. It was difficult to detect high-frequency, low-amplitude waves in that recording. Also clinicians rarely needed to study the EEG frequencies of more than 20 Hz. Study of high frequency (>25 Hz) EEG waves became easier with the introduction of digital signal recording and processing techniques. In the late 1970s, electronic computers started being used in EEG processing and in the 1980s came the concept of the *binding problem* (see Chapter 5) in cognitive psychology research. Proponents of the binding problem believe that different features of the same entity (such as voice and look of a single individual), processed in different parts of the brain need to be integrated in order to form the concept of that entity in our mind as a whole. They also believe that this happens through simultaneous firing of neurons in different parts of our brain, without any structural connections, in the gamma frequency range. Although gamma is from 30 to 80 Hz, 40 Hz is considered to be of somewhat representative importance from the cognition point of view (Jefferys et al., 1996). For the computational analysis purpose, taking the band from 37 to 43 Hz might be useful (Miltner et al., 1999).

Gamma rhythms occur in humans and other mammals following sensory stimuli, often in brief runs. Gamma rhythms have been identified in all parts of the sensory cortex and also in the hippocampus. Apart from sensory information processing, and perception gamma EEG has been implicated in the formation of memory, linguistic processing, internal thoughts (apparently without any outside stimuli), and behavior, particularly motor actions and planning for the actions. It has also been implicated in attention, arousal, and object recognition.

Not all researchers believe in the importance of the EEG gamma wave. According to one group, the gamma wave might be merely due to muscle activities, that is, it is EMG rather than EEG from the brain (Whitham et al., 2007, 2008).

Apart from the frequency bands discussed above there are EEG frequency bands of interest beyond 80 Hz. They are called high-gamma (80–150 Hz) bands (Canolty et al., 2006). Recently a new EEG component called the phi complex (9.2–11.5 Hz), located above the right centroparietal cortex, has been implicated in social coordination tasks (Tognoli et al., 2007).

References

Bailet, S., J. C. Mosher, and R. M. Leahy, Electromagnetic brain mapping, *IEEE Signal Proc. Mag.*, **18(6)**: 14–30, 2001.

Banquet, J., Spectral analysis of the EEG in meditation, *Electroencephal. Clin. Neurophysiol.*, **35**: 143–151, 1973.

Benbadis, S. R., Introduction to sleep electroencephalography, in: T. L. Lee-Cheong (ed.), *Sleep: A Comprehensive Handbook*, John Wiley & Sons, Hoboken, NJ, 2006.

Benbadis, S. R. and D. Rielo, EEG artifacts, *eMedicine*, 2008 available at http://emedicine.medscape.com/article/1140247-overview.

Benchenane, K., A. Peyrache, M. Khamassi, P. L. Tierney, Y. Gioanni, F. P. Battaglia, and S. I. Wiener, Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning, *Neuron*, **66**: 921–936, 2010.

Bentivoglio, M., L. Tassi, E. Pech, C. Costa, P. F. Fabene, and R. Spreafico, Cortical development and focal cortical dysplasia, *Epileptic Disord.*, **5**: S27–S34, 2003.

Borcea, L., Electrical impedance tomography: Tropical review, *Inverse Probl.*, **18**: R99–R136, 2002.

Buzsaki, G., Theta oscillations in the hippocampus, *Neuron*, **33**: 325–340, 2002.

Buzsaki, G., Neuronal oscillations in cortical networks, *Science*, **304**: 1926–1929, 2004.

Canolty, R. T., E. Edwards, S. S. Dalal, M. Soltani, S. S. Nagarajan, H. E. Kirsch, M. S. Berger, N. M. Barbaro, and R. T. Knight, High gamma power is phase-locked to theta oscillations in human neocortex, *Science*, **313**: 1626–1628, 2006.

Cantero, J. L., M. Atienza, R. Stickgold, M. J. Kahana, J. R. Madsen, and B. Kocsis, Sleep-dependent θ oscillations in human hippocampus and neocortex, *J. Neurosci.*, **23**: 10897–10903, 2003.

Delorme, A. and S. Makeig, EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, *J. Neurosci. Methods*, **134**: 9–21, 2004.

EEGLAB, 2017, http://sccn.ucsd.edu/eeglab/.

Fischl, B. and A. M. Dale, Measuring the thickness of the human cerebral cortex from magnetic resonance imaging, *Proc. Natl. Acad. Sci. USA*, **97(20)**: 11050–11055, 2000.

Hallez, H., B. Venrumste, R. Grech, J. Muscat, W. De Clercq, A. Vergult, Y. D'Asseler et al., Review on solving the forward problem in EEG source analysis, *J. Neuroeng. Rehabil.*, **4**: 46, 2007 available at http://www.jneuroengrehab.com/content/4/1/46.

Harmony, T., T. Fernandez, J. Silva, J. Bernal, L. Diaz-Comas, A. Reyes, E. Marosi, M. Rodriguez, and M. Rodriguez, EEG delta activity: An indicator of attention to internal processing during performance of mental tasks, *Int. J. Psychophsiol.*, **24**: 161–171, 1996.

Itil, T. M., Quantitative and qualitative EEG findings in schizophrenia, *Schizophrenia Bull.*, **3(1)**: 61–79, 1977.

Jefferys, J. G. R., R. D. Traub, and M. A. Whittington, Neuronal networks for induced '40 Hz' rhythms, *Trends Neurosci.*, **19**: 202–208, 1996.

Joyce, C. A., I. F. Gorodnitsky, and M. Kutas, Automatic removal of eye movement and blink artifacts from EEG data using blind component separation, *Psychophysiology*, **41**: 1–13, 2004.

Kandel, E. R., J. H. Schwartz, and T. M. Jessel, *Principles of Neural Science*, 4th ed., McGraw Hill, New York, 2000.

Keren, A. S., S. Yuval-Greenberg, and L. Y. Deouell, Saccadic spike potential in gamma-band EEG: Characterization, detection and suppression, *NeuroImage*, **49**: 2248–2263, 2010.

Klimesch, W., EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis, *Brain Res. Rev.*, **29**: 169–195, 1999.

Marshall, L., H. Helgadottir, M. Molle, and J. Born, Boosting slow oscillations during sleep potentiates memory, *Nature*, **444**: 610–613, 2006.

Michel, C. M., M. Koukku, and D. Lehmann, EEG reactivity in high and low symptomatic schizophrenics, using source modeling in frequency domain, *Brain Topogr.*, **5(4)**: 389–394, 1993.

Miltner, W. H. R., C. Braun, M. Arnold, H. Witte, and E. Traub, Coherence of gamma-band EEG activity as a basis for associative learning, *Nature*, **397**: 434–436, 1999.

Murakami, S. and Y. Okada, Contributions of principal neocortical neurons to magnetoencep-halography and electroencephalography signals, *J. Physiol.*, **575(3)**: 925–936, 2006.

Niedermeyer, E., Alpha rhythms as physiological and abnormal phenomena, *Int. J. Psychophysiol.*, **26**: 31–49, 1997.

Olejniczak, P., Neurophysiologic basis of EEG, *J. Clin. Neurophysiol.*, **23(3)**: 186–189, 2006.

Oostenveld, R. and P. Praamstra, The five percent electrode system for high-resolution EEG and ERP measurements, *Clin. Neurophysiol.*, **112**: 713–719, 2001.

Perlis, M. L., H. Merica, M. T. Smith, and D. E. Giles, Beta EEG activity and insomnia, *Sleep Med. Rev.*, **5(5)**: 365–376, 2001.

Pfurtscheller, G., C. Brunner, A. Schlogl, and F. H. Lopes da Silva, Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks, *NeuroImage*, **31**: 153–159, 2006.

Pfurtscheller, G. and F. H. Lopes da Silva, Event-related EEG/MEG synchronization and desynchronization: Basic principles, *Clin. Neurophysiol.*, **110**: 1842–1857, 1999.

Ray, W. J. and H. W. Cole, EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes, *Science*, **228**: 750–752, 1985.

Sekimoto, M., M. Kato, T. Watanabe, N. Kajimura, and K. Takahashi, Cortical regional differences of delta waves during all-night sleep in schizophrenia, *Schizophrenia Res.*, **126**: 284–290, 2011.

Tognoli, E., J. Lagarde, G. C. DeGuzman, and J. A. Scott Kelso, The phi complex as a neuromarker of human social coordination, *Proc. Natl. Acad. Sci. USA*, **104(9)**: 8190–8195, 2007.

Whitham, E. M., T. Lewis, K. J. Pope, S. P. Fitzgibbon, C. R. Clark, S. Loveless, D. DeLosAngeles, A. K. Wallace, M. Broberg, and J. O. Willoughby, Thinking activates EMG in scalp electrical recordings, *Clin. Neurphysiol.*, **119(5)**: 1166–1175, 2008.

Whitham, E. M., K. J. Pope, S. P. Fitzgibbon, T. Lewis, C. R. Clark, S. Loveless, M. Broberg et al., Scalp electrical recording during paralysis: Quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG, *Clin. Neurphysiol.*, **118(8)**: 1877–1888, 2007.

2

Preprocessing

We have already seen in Chapter 1 that EEG recorded from the scalp often comes mixed up with signals other than those from the brain. But in EEG processing we are only interested in what is happening inside the brain. For this, signals from sources other than the brain need to be removed as far as possible from the recorded EEG, before undertaking the actual processing. The various processes involved in this removal are generically termed as *preprocessing*. Even in this age of sophisticated computer hardware and software the gold standard of preprocessing remains the removal of artifacts by visual inspection. To aid the removal by visual inspection special recordings are also performed, such as electrooculogram (EOG) and ECG (Figure 1.9). They track the eye movement and heartbeat, respectively. In this book we will not cover preprocessing by visual inspection. Often, channels and trials with an undesirable amount of artifacts are removed altogether from the EEG to be processed. Computer algorithms have also been developed for automatic preprocessing. We can classify them broadly into two categories: (1) filtering and (2) decomposition either by independent component analysis (ICA) or principal component analysis (PCA). In the next two sections, we will discuss about them. One important thing to remember is that conventional filtering (such as impulse response filtering) removes the artifacts, whereas both PCA and ICA can only identify the artifacts but cannot remove them. In order to remove the artifacts from the data we need separate techniques.

2.1 Filtering

The most widely used filter on the EEG is the band-pass filter. The band-pass filter is a combination of a low-pass filter and a high-pass filter. Let $s(t)$ be an EEG signal. $s\left(t_i\right)_{i=1}^{N}$ is the digitized signal with N discrete points. The time interval between any two successive points is always fixed. Then we can simply write $s\left(i\right)_{i=1}^{N}$. $s(i)$ is called an *impulse* or *amplitude* at time i. We will write $s\left(i\right)_{i=1}^{N}$ as $s(i)$ where there is no risk of confusion. Digital filters are designed by *impulse response*. If there are only a finite number of impulse responses we call it a *finite impulse response* (FIR) filter, and if there are an infinite number of impulse responses we call it an *infinite response* (IIR) filter. An impulse

response to a signal $s(i)$ is just another (artificial) signal $h(i)$. If $h(i)$ has only a finite number of nonzero points it is called a finite response; otherwise it is called an infinite response. The filtered output signal $y(i)$ is expressed as

$$y(i) = s(i) * h(i), \tag{2.1}$$

where $*$ denotes a convolution operation. To understand the theoretical principles of filter design knowledge of complex analysis (e.g., see Ahlfors, 1979) is advantageous. However, here we will give a brief conceptual idea of digital filter design without resorting to complex analysis as far as possible.

The design of an impulse response filter (the most commonly used filter type) is all about finding the right $h(i)$ in (2.1). However, since (2.1) is in the time domain and filters are generally frequency specific, we need to find suitable transformations for each of $y(i)$, $s(i)$, and $h(i)$ so that an appropriately transformed 'copy' of $h(i)$ can be obtained (designing the filter in the transformed space becomes much more convenient). Then, by performing an appropriate inverse transform on that copy, the right $h(i)$ can be found. This 'suitable transform' turns out to be the z-transform, which is a generalization of the discrete Fourier transform (DFT) and maps the time domain signals into the complex plane. The time domain signals can be recovered by the inverse z-transform. The z-transformation $S(z)$ of a signal $s(i)$ is defined as

$$S(z) = \sum_{i=-\infty}^{\infty} s(i) z^{-i}. \tag{2.2}$$

An important property of z-transformation is that if we apply it on both sides of (2.1) we will get

$$Y(z) = S(z) H(z), \tag{2.3}$$

where $Y(z)$ and $H(z)$ are the z-transforms of $y(i)$ and $h(i)$, respectively. The advantage of (2.3) is that a relatively cumbersome convolution operation in (2.1) is replaced by a simpler multiplication operation in (2.3). Also Y, S, H are complex functions, which are more convenient to deal with mathematically than a discrete time series like y, s, h.

We already know that designing a filter is all about finding an appropriate $h(i)$ in (2.1), which, under the z-transformation, reduces to finding an appropriate $H(z)$ in (2.3). We will see that finding a complex function $H(z)$ will be more convenient than finding $h(i)$ directly. Once $H(z)$ is found, $h(i)$ can be obtained from the inverse z-transformation as defined below.

$$h(i) = \frac{1}{2\pi j} \oint_C H(z) z^{i-1} dz, \tag{2.4}$$

where $j = \sqrt{-1}$, C is a closed curve lying within the region of convergence of $H(z)$. For the details of z-transform see chapter 3 of Proakis and Manolakis (2007).

2.1.1 Impulse Response Filter

Suppose, in an EEG signal, we need to study the gamma band (30–80 Hz). We typically pass the signal between 30 and 80 Hz through the band-pass filter (see Section 2.1.4). This band-pass filter consists of one high-pass filter, which allows all frequencies greater than equal to 30 Hz to pass through and suppresses all frequencies below 30 Hz (it diminishes the amplitudes associated with frequencies less than 30 Hz to a great extent to suppress the part of the signal with frequencies below 30 Hz) and one low-pass filter, which allows all frequencies less than equal to 80 Hz to pass through and suppresses all frequencies above 80 Hz (it diminishes the amplitudes associated with frequencies greater than 80 Hz to a great extent to suppress the part of the signal with frequencies above 80 Hz).

Figure 2.1 elaborates how a low-pass filter can be implemented. It depicts the graph of the complex function $H(z)$, where $z = F$, F is the normalized frequency. It is about an FIR filter. There are only a finite number of time points i, such that $h(i) \neq 0$. The total of the frequencies is equal to the sample frequency. The normalized frequency is obtained by dividing the actual frequency by the sample frequency (multiplied by π radian/sample), where 0.5 denotes the Nyquist frequency.

If we write $z = \exp(j\omega_k)$ in $H(z)$, $\omega_k = 2\pi k/N$, we get the *frequency response* of the digital filter. We prefer to write $H(\exp(j\omega_k))$ as $H(\omega_k)$. $|H(\omega_k)|$ gives the

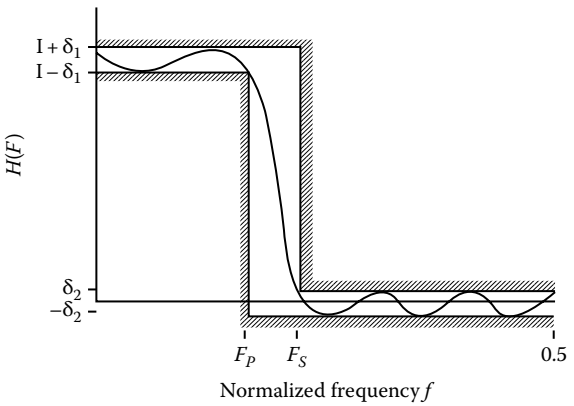


FIGURE 2.1
The graph of $H(z)$, where z has been replaced by normalized frequency. δ_1 is the allowed passband deviation, δ_2 the allowed stopband deviation, F_P the passband edge, and F_S the stopband edge. (Taken from Parks, T.W. and McClelland, J.H., *IEEE Trans. Circuit Theory*, 19(2), 189, 1972.)

amplitude or the *magnitude response*. If a sinusoidal signal $A\sin(2\pi n\,t+\alpha_n)$ is passed through a filter, in general, it will become $B\sin(2\pi n\,t+\beta_n)$. By *phase response* we understand $\beta_n-\alpha_n$.

In the design of frequency selective filters, the desired filter characteristics are specified in the frequency domain in terms of the desired magnitude and phase response of the filter. The issue of the type of filter to design, FIR or IIR, depends on the nature of the problem and on the specifications of the desired frequency response.

In practice, FIR filters are employed in filtering problems where there is a requirement for a linear phase characteristic within the passband of the filter (the graph of $\beta_n-\alpha_n$ against n is a straight line, where n varies across the passband). Under certain conditions an FIR filter can always have the linear phase response. If there is no requirement for a linear phase characteristic, either an IIR or an FIR filter may be employed. However, as a general rule, an IIR filter has lower side lobes in the stopband than an FIR filter having the same number of parameters. For this reason, if some phase distortion is either tolerable or unimportant, an IIR filter is preferable, primarily because its implementation involves fewer parameters, requires less memory, and has lower computational complexity (Proakis and Manolakis, 2007, p. 654).

For EEG signals, phase information is important in some cases, such as phase synchronization measure. So we need to have access to both FIR and IIR filter design methodologies. In EEG processing, typically, *frequency-sampling design* of FIR filters is followed. In this method we sample the desired frequency response $H(\omega)$, where $\omega\in\{\omega_k\}_{k=0}^{N-1}$, as shown in Figure 2.2. Note that ω is the *circular frequency* or the *angular velocity* expressed as $\omega=2\pi/T$, where T is the *period* given by $T=1/f$, where f is the normalized linear frequency. In the case of a filter $f=f_C/f_S$, where f_C is the cutoff frequency and f_S is the sampling frequency. The unit of ω is radians per second, whereas unit of f is Hz.

In the case of an FIR signal (2.1) can be written as

$$y(i)=\sum_{k=0}^{N-1}h(k)s(i-k). \tag{2.5}$$

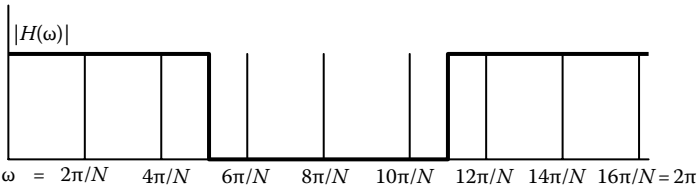


FIGURE 2.2
Amplitude response $|H(\omega)|$ of the FIR filter (thick line) is plotted against the circular frequency ω.

$h(k)$'s are the *filter coefficients* (also known as *taps*) and N is the *filter order*. N equally spaced points between 0 and 2π have to be taken as shown in Figure 2.2. $\omega_k = 2\pi k/N, k = 0, 1, \ldots, N-1$. For odd N $H(0)$ will have to be real and

$$H(N-k) = \overline{H(k)} \quad \text{for } k = 1, \ldots, \frac{N-1}{2}, \tag{2.6}$$

where $\overline{H(k)}$ is the complex conjugate of $H(k)$. For even N $H(0)$ will have to be real, and

$$H(N-k) = \overline{H(k)} \quad \text{for } k = 1, \ldots, \frac{N}{2} - 1 \text{ and } H\left(\frac{N}{2}\right) = 0. \tag{2.7}$$

Conditions (2.6) and (2.7) are imposed to make the filter with real coefficients $h(k)$ in (2.5). The $h(k)$'s are simply obtained by inverse discrete Fourier transformation (iDFT) of $H(k)$. That is,

$$h(n) = \frac{1}{N} \sum_{k=0}^{N-1} H(k) \exp\left(j2\pi \frac{n}{N} k\right) \quad \text{for } n = 0, 1, \ldots, N-1. \tag{2.8}$$

If condition (2.6) holds, then (2.8) becomes

$$h(n) = \frac{H(0)}{N} + \frac{2}{N} \sum_{k=1}^{(N-1)/2} \mathrm{Re}\left\{H(k) \exp\left(j2\pi \frac{n}{N} k\right)\right\}. \tag{2.9}$$

If condition (2.7) holds, then (2.8) becomes

$$h(n) = \frac{H(0)}{N} + \frac{2}{N} \sum_{k=1}^{(N/2)-1} \mathrm{Re}\left\{H(k) \exp\left(j2\pi \frac{n}{N} k\right)\right\}. \tag{2.10}$$

Let us now apply the method to the following FIR filter design problem.

Example 2.1 (Rao and Gejji, 2010): Design a 17-tap linear phase low-pass FIR filter with a cutoff frequency $\omega = \pi/2$ radian/s.

Solution: The ideal low-pass frequency response with a linear phase is

$$H(\omega) = \begin{cases} \exp\left(-j\dfrac{(N-1)\omega}{2}\right), & 0 < \omega \le \dfrac{\pi}{2} \\ 0, & \dfrac{\pi}{2} < \omega \le \pi \end{cases}. \tag{2.11}$$

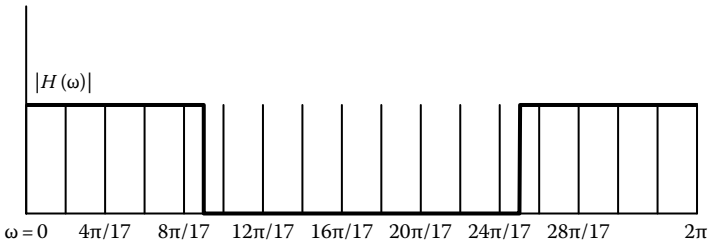


FIGURE 2.3
Amplitude response of the 17-tap FIR filter (thick line).

The vertical lines in Figure 2.3 indicate the frequency sampling taken at $\omega = 2\pi k/17$, for $k = 0, 1, \ldots, N-1$. From Figure 2.3 we find that

$$|H(k)| = \begin{cases} 1, & 0 \le k \le 4 \\ 0, & 5 \le k \le 12 \\ 1, & 13 \le k \le 16 \end{cases}. \tag{2.12}$$

The phase response of the filter is shown in Figure 2.4. Putting $N = 17$ in (2.11) we get

$$H(\omega) = \begin{cases} \exp(-j8\omega), & 0 \le \omega \le \dfrac{\pi}{2} \\ 0, & \dfrac{\pi}{2} < \omega \le \pi \end{cases}. \tag{2.13}$$

From (2.13) we get the phase θ_k of the filter to be

$$\begin{aligned} \theta_k &= -8\omega_k = -8 \cdot \frac{2\pi k}{N}, & 0 \le k \le 8 \\ &= -\frac{16\pi}{17}(k - 17), & 9 \le k \le 16. \end{aligned} \tag{2.14}$$

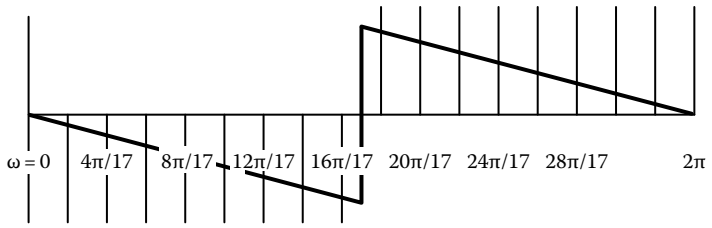


FIGURE 2.4
Phase response of the 17-tap FIR filter (thick line), with respect to the frequency.

Since $H(k)$ is complex, we may write $H(k) = |H(k)| \exp(j\theta_k)$, which, by (2.12), becomes

$$H(k) = \begin{cases} \exp\left(-\dfrac{16\pi k}{17}\right), & 0 \le k \le 4 \\ 0, & 5 \le k \le 12 \\ \exp\left(-\dfrac{16\pi}{17}(k-17)\right), & 13 \le k \le 16 \end{cases}. \tag{2.15}$$

We apply iDFT of Equation 2.9 on $H(k)$ to get back the $h(n)$ of (2.9). That is,

$$\begin{aligned} h(n) &= \frac{1}{N}\left[H(0) + 2\sum_{k=1}^{(N-1)/2} \mathrm{Re}\left\{H(k)\exp\left(j\frac{2\pi nk}{N}\right)\right\}\right] \\ &= \frac{1}{17}\left[1 + 2\cos\left\{\frac{2\pi}{17}(n-8)\right\} + 2\cos\left\{\frac{4\pi}{17}(n-8)\right\}\right. \\ &\quad \left. + 2\cos\left\{\frac{6\pi}{17}(n-8)\right\} + 2\left\{\frac{8\pi}{17}(n-8)\right\}\right], \quad 0 \le n \le 16. \end{aligned} \tag{2.16}$$

The FIR filter coefficients are tabulated below:

n	*h*(*n*)	*n*	*h*(*n*)	*n*	*h*(*n*)
0	0.0398	6	−0.0299	12	0.03154
1	−0.0488	7	0.31876	13	0.06598
2	−0.03459	8	0.5294	14	−0.03459
3	0.06598	9	0.31876	15	−0.0488
4	0.03154	10	−0.0299	16	0.0398
5	−0.10747	11	−0.10747		

This example is just to give a feel of how FIR filters are designed. More examples can be found in Rao and Gejji (2010). A collection of good software is available for filter design. One such is the Signal Processing Toolbox in MATLAB®. In this graphical user interface (GUI)–based software the cutoff frequency, number of taps, window type, etc., are specified in appropriate boxes. After generation of the filter, the code can also be imported into the working directory for application on the signals.

2.1.2 Butterworth Low-Pass Filter

Digital FIR filters are designed with a finite number of discrete impulse responses as shown in Example 2.1, and this is a direct process. In contrast, IIR filters are designed first for the analog case and then discretized to get the digital IIR filters. This is convenient, because a large library of analog IIR filters is already available. The oldest and the most widely used analog IIR filters are the Chebyshev filter and the Butterworth filter. Among these two, the Butterworth filter is more popular for the EEG processing. The major reasons are elaborated in Figure 2.5.

As we have already mentioned, in many applications, preserving the phase information in the EEG is desirable, and therefore a filter with a linear or almost linear phase response is preferred. Also, the original voltage in the scalp EEG is only a few tens of microvolts, which is very faint and can therefore be easily contaminated due to variation introduced by the filter in terms of spurious frequencies, and therefore a filter with a more uniform passband and stopband amplitude response is preferred. These are the main reasons for the Butterworth filter being preferred for EEG processing as demonstrated in Figure 2.5. However, one drawback of the Butterworth filter is that its transition from the passband to the stopband is more gradual (see Figure 2.6), which is sharper in the Chebyshev filter.

The Butterworth filter is characterized by its magnitude response function as

$$\left|H(\omega)\right| = \frac{1}{\left\{1+\left(\frac{\omega}{\omega_C}\right)^{2N}\right\}^{\frac{1}{2}}}, \tag{2.17}$$

where

ω_C is the circular cutoff frequency
N is the order of the filter

We can design a low-pass Butterworth filter to meet the following frequency domain specifications.

$$K_P \le 20\log\left|H(\omega)\right| \le 0 \quad \text{for all } \omega \le \omega_P,$$

and

$$20\log\left|H(\omega)\right| \le K_S \quad \text{for all } \omega \ge \omega_{ST}, \tag{2.18}$$

where

K_P and K_S are passband gain and stopband gain, respectively
ω_P is the circular passband frequency
ω_{ST} is the circular stopband frequency

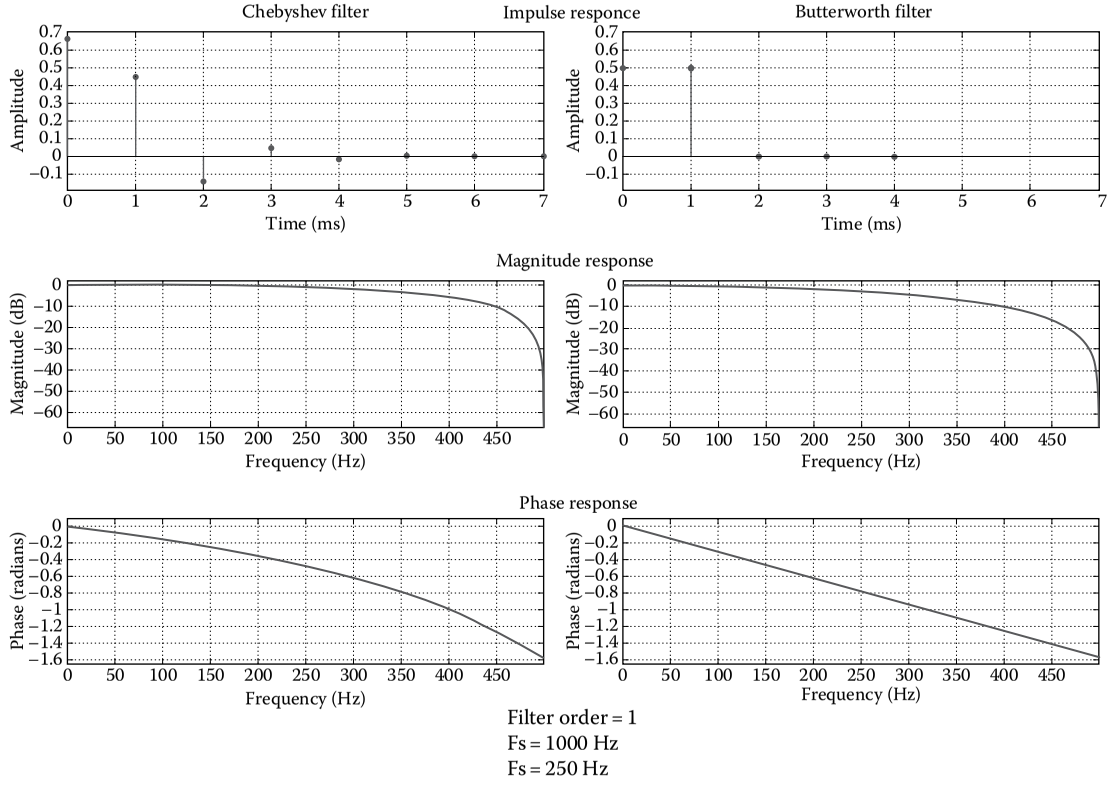


FIGURE 2.5
Impulse response, magnitude response, and phase response of Butterworth and Chebyshev low-pass filters with the same parameter values. Notice the more uniform impulse response both in the passband and in the stopband of the Butterworth filter than the Chebyshev filter. The phase response of the Butterworth filter is more linear than that of the Chebyshev filter.

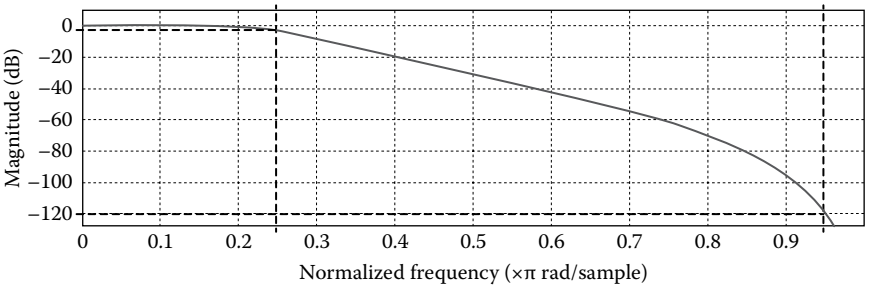


FIGURE 2.6
Amplitude-response graph of a low-pass Butterworth filter with cutoff frequency $\pi/4$ radian/s. The left-most vertical dashed line denotes ω_C, the right-most vertical dashed line denotes ω_{ST}, the uppermost horizontal dashed line denotes K_P, and the lowermost horizontal dashed line denotes K_S.

The geometric significance of these parameters in the amplitude response diagram is shown in Figure 2.5.

Taking 20log on both sides of (2.17) we get

$$20\log\left|H(\omega)\right| = K = -10\log\left\{1+\left(\frac{\omega}{\omega_C}\right)^{2N}\right\}. \tag{2.19}$$

When $K = K_P$, $\omega = \omega_P$, where ω_P is the passband frequency. So $K_P = -10\log\left\{1+\left(\frac{\omega_P}{\omega_C}\right)^{2N}\right\}$ or $\left(\frac{\omega_P}{\omega_C}\right)^{2N} = 10^{-\frac{K_P}{10}} - 1$. Similarly, we get $\left(\frac{\omega_{ST}}{\omega_C}\right)^{2N} = 10^{-\frac{K_S}{10}} - 1$. By eliminating ω_C from the last two expressions and simplifying, we get

$$N = \frac{\log\left\{\frac{10^{-\frac{K_P}{10}}-1}{10^{-\frac{K_S}{10}}-1}\right\}}{2\log\left(\frac{\omega_P}{\omega_{ST}}\right)}. \tag{2.20}$$

After the order of the filter is fixed the cutoff frequency can be fixed in either of the two ways (from the same two expressions):

$$\omega_C = \frac{\omega_P}{\left(10^{-\frac{K_P}{10}}-1\right)^{\frac{1}{2N}}}, \quad \text{when } K_P \text{ is known.} \tag{2.21}$$

$$\omega_C = \frac{\omega_{ST}}{\left(10^{-\frac{K_S}{10}} - 1\right)^{\frac{1}{2N}}}, \quad \text{when } K_S \text{ is known.} \tag{2.22}$$

Also, ω_C can be determined by averaging the two values obtained by (2.21) and (2.22).

2.1.3 Gaussian Low-Pass Filter

A Gaussian filter is one whose impulse response is a Gaussian function. When a Gaussian filter is used, more weights are assigned to the particular point where Gaussian filtering is being applied and higher weights closer to the same point, whereas less and less weights are assigned to the points away from the point of application of the filter. This width of influence is controlled by the variance of the Gaussian. This may be particularly useful in preserving some of the distinct shapes of the EEG, such as epileptic seizure spikes and sharp waves over the spikes of artifacts.

In the case of IIR/FIR, both feedforward and feedback signal points are considered. That is, for applying IIR/FIR at a particular point (say, the nth point), for the signal values of the previous points ($n-1$, $n-2$, etc., up to the length of the IIR/FIR window), one also has to consider the filtered signal (that is output from $n-1$, $n-2$, etc.). This second part is the major difference of the IIR/FIR filters with the Gaussian filter. However, it is quite possible to approximate Gaussian filtering with some variations of IIR/FIR filters, although the implementation may become slightly more complicated (Majumdar and Vardhan, 2011). The impulse response function in the analog form is given by

$$h(t) = \alpha f_C \exp\left\{-\pi\left(\alpha f_C t\right)^2\right\}, \tag{2.23}$$

where

α is a constant

f_C is the linear cutoff frequency

For a digital version just replace the continuous time t by discrete time i. For more details on the Gaussian filter, see Raja et al. (2002).

Here is a MATLAB code for the implementation of the Gaussian low-pass filter. Like most other IIR filters the Gaussian filter was also originally designed for analog signals. Later, its digital form was obtained by digitizing the analog filter. In this implementation the Gaussian low-pass filter has been implemented as an FIR filter (after digitizing and truncating).

```
function b=gaussfiltcoef(SR, fco)
%GAUSSFILTCOEF Return coefficients of Gaussian lowpass filter.
% SR=sampling rate, fco=cutoff (-3dB) freq, both in Hz.
% Coeffs for FIR filter of length L (L always odd) are computed.
% This symmetric FIR filter of length L=2N+1 has delay N/SR
  seconds.
% Examples of use
%  Compute Gaussian filter frequency response for SR=1000,
   fco=50 Hz:
%  freqz(gaussfiltcoef(1000,50),1,256,1000);
%  Filter signal X sampled at 5kHz with Gaussian filter with
   fco=500:
%  y=filter(gaussfiltcoef(5000,500),1,X);
%  SR, fco are not sanity-checked. WCR 2006-10-11.
b=0;
a=3.011*fco;
N=ceil(0.398*SR/fco); %filter half-width, excluding midpoint
%Width N corresponds to at least +-3 sigma which captures at
%least 99.75% of area under Normal density function.
%sigma=1/(a*sqrt(2pi)).
L=2*N+1;          %full length of FIR filter
for k=-N:N
b(k+N+1)=3.011*(fco/SR)*exp(-pi*(a*k/SR)^2);
end;
%b(k) coeffs computed above will add to almost exactly unity,
%but not quite exact due to finite sampling and truncation
%at +- 3 sigma.
%Next line adjusts to make coeffs b(k) sum to exactly unity.
b=b/sum(b);
```

2.1.4 Band-Pass Filter

In the last section, we discussed about the low-pass filter. In this section, we touch upon the high-pass filter, which is very similar to the low-pass filter both in theory and in implementation. In fact, any low-pass filter can be transformed into a high-pass filter (see Proakis and Manolakis, 2007), and then, by combining a low-pass filter with a high-pass filter, a band-pass filter (Figure 2.7) can be created. The simplest way to transform a low-pass filter into a high-pass one is

$$H_{hp}(\omega) = H_{lp}(\omega - \pi), \tag{2.24}$$

where

$H_{hp}(\omega)$ is the frequency response function of the high-pass filter
$H_{lp}(\omega)$ is that of the low-pass filter

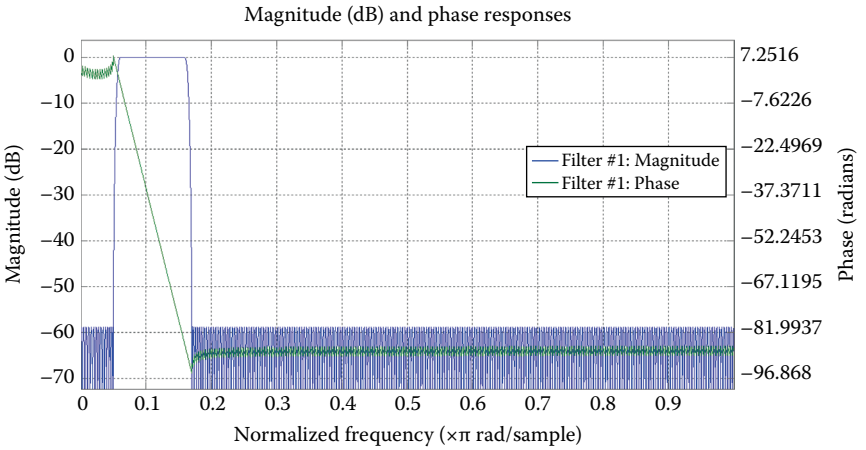


FIGURE 2.7
Normalized frequency (×π radian/sample) versus magnitude (in dB) plot. Magnitude (in blue) and phase (in green) response of a bandpass FIR filter. Passband is 30–80 Hz. Sample frequency is 1000 Hz. Stopband attenuation is 60 dB.

Since a frequency response of π radians is equivalent to multiplication of the impulse response (of the low-pass filter) $h_{lp}(i)$ by $\exp(j\pi n)$, the impulse response of the high-pass filter is

$$h_{hp}(i) = \left(\exp(j\pi)\right)^i h_{lp}(i) = (-1)^i h_{lp}(i). \tag{2.25}$$

Therefore, the impulse response of the high-pass filter is simply obtained from the impulse response of the low-pass filter by changing the sign of the odd-numbered impulses in $h_{lp}(i)$.

Conversely,

$$h_{lp}(i) = (-1)^i h_{hp}(i). \tag{2.26}$$

In the implementation of a band-pass filter in a popular scientific software like MATLAB, a chosen low-pass filter is transformed into the corresponding high-pass filter (Proakis and Manolakis, 2007, p. 334). Then combining both a band-pass filter is created. In this sense, a Butterworth or Chebyshev band-pass filter is created and implemented in MATLAB.

The opposite of a band-pass filter is a *band-stop* filter, which stops a particular frequency band and allows all other frequencies to pass through. In EEG processing, a band-stop filter is used typically to filter out the electrical line noise from the data. The main alternating current (AC) electric supply comes either at the 50 Hz cycle or at the 60 Hz cycle, depending on the country. If the 50 Hz line noise is to be eliminated, often a band-stop filter

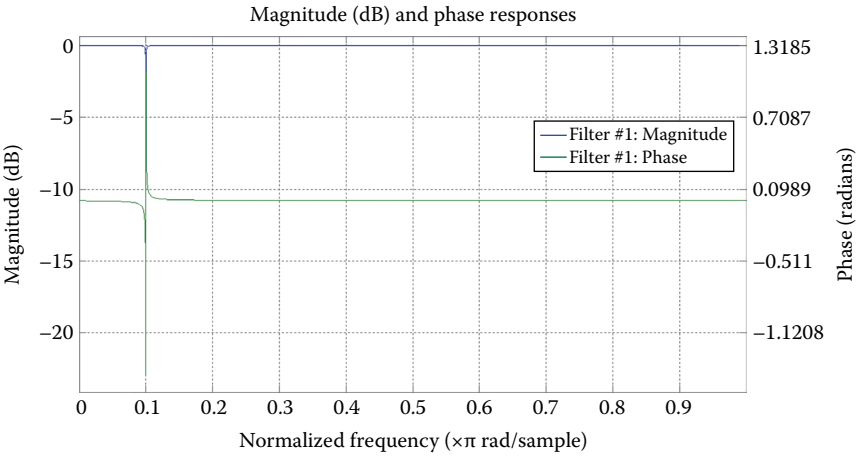


FIGURE 2.8
Magnitude (blue curve) and phase response (green curve) of a band-stop or notch filter with a stopband at 50 Hz.

that stops the frequency band from 49 to 51 Hz (or 48 to 52 Hz, depending on the implementation) is used (Figure 2.8).

2.2 Decomposition Techniques

Conventional signal filtering methods eliminate certain frequency bands from the EEG signals, irrespective of whether they are coming from the brain or from somewhere outside. So, along with noise or artifacts these methods are likely to eliminate useful information too, if that happens to fall within the eliminated frequency bands. In principle, sources of brain signals and the sources of noise or artifacts are generally independent of each other. So, if the collected raw signals can be decomposed into the source signals, noise can be separated out and eliminated, while useful information can be retained intact even if they are in the same frequency range as the noise. This notion has given rise to two powerful decomposition techniques. One is called principal component analysis (PCA), which is more restricted, and the other is called independent component analysis (ICA), which is more general in its scope of applications. The PCA and ICA techniques have achieved great success in eliminating EEG artifacts (Delorme and Makeig, 2004).

2.2.1 Principal Component Analysis

Suppose the EEG is collected through n channels and they are generated by m sources, where $m < n$ (in many neuroscience applications $m \gg n$, but if we

consider only the 'dominant' or 'principal' sources it may so happen that $m < n$). PCA will help us identify these m independent sources out of the n signals. First consider a window p time points long. Consider the matrix of n rows and p columns. This is a data matrix $A_{n\times p}$.

Step 1: Find the mean value of each row and subtract it from all entries of that particular row. Each row becomes a normalized data vector with zero mean. Let us denote it by $A^N_{n\times p}$. Here superscript N stands for normalized.

Step 2: Now form the covariance matrix $C_{n\times n}$, whose ijth entry is the covariance between the ith and the jth rows (EEG channels) of $A_{n\times p}$.

Step 3: Find eigenvalues of the covariance matrix $C_{n\times n}$.

Step 4: Place the eigenvalues in descending order.

Step 5: Find the eigenvector corresponding to each eigenvalue. The eigenvector associated with the highest eigenvalue is called the *principal component*.

Step 6: Form the matrix E with rows as the eigenvector. $EA^N_{n\times p}$ will give the new transformed data that is decomposed along its constituent components.

In our original example, m eigenvalues will have 'large' nonzero values. The remaining $n-m$ eigenvalues will have values close to zero. These $n-m$ small eigenvalues may be ignored with little consequence on the final outcome. If out of these m components some are due to artifacts, such as eyeball movements, after PCA decomposition in the $EA^N_{n\times p}$, they will be more easily discernable than in the raw data. To remove the artifacts additional elimination schemes are required.

Here is a MATLAB code due to Shlens (2009) for implementing the above PCA algorithm. The author's original notations in the code have been retained and therefore they are different from the notations used in the text.

```
function [signals,PC,V] = pca1(data)
% PCA1: Perform PCA using covariance.
% data - MxN matrix of input data
% signals - MxN matrix of projected data
% PC - each column is a PC
% V - Mx1 matrix of variances
[M,N] = size(data);
% subtract off the mean for each dimension
mn = mean(data,2);
data = data - repmat(mn,1,N);
% calculate the covariance matrix
covariance = 1 / (N-1) * data * data';
% find the eigenvectors and eigenvalues
[PC, V] = eig(covariance);
```

```
% extract diagonal of matrix as vector
V = diag(V);
% sort the variances in decreasing order
[junk, rindices] = sort(-1*V);
V = V(rindices);
PC = PC(:,rindices);
% project the original data set
signals = PC' * data;
```

Here, we have given a very brief description of the PCA technique. For a comprehensive coverage of PCA the readers are encouraged to consult Joliffe (2002). Two excellent online tutorials by Smith (2002) and Shlens (2009) are worth mentioning.

2.2.2 Independent Component Analysis

ICA is a generalization of the PCA. When the independent components are mutually orthogonal ICA becomes PCA. Since orthogonality is a strong condition, ICA usually has a wider application than PCA. The central idea of ICA is the so-called blind source separation (BSS), which can be formulated as

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.27}$$

where

- $\mathbf{x}$ is a vector of m sensors (by default, all vectors in this book are column vectors unless mentioned otherwise)
- $\mathbf{s}$ is a vector of n (independent) sources
- $\mathbf{A}$ is an $m \times n$ matrix, called the *mixing matrix*

Each component of $\mathbf{x}$ and $\mathbf{s}$ is a time series data. A more generalized form of (2.27) is

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}, \tag{2.28}$$

where $\mathbf{n}$ is additive noise, which is often modeled as Gaussian white noise. In actual data processing, before applying the ICA preprocessing is done (such as removing exceptionally noisy channels by visual inspection) in order to eliminate $\mathbf{n}$ from (2.28) as far as possible.

In real data, $\mathbf{A}$, $\mathbf{s}$ and n, that is, the number of sources, are unknown, which needs to be found out. The very first step is to make the mean of $\mathbf{x}$ zero. This can easily be achieved by subtracting the average of the data from itself. Next, each component of $\mathbf{x}$ is divided by its standard deviation, in order to make the variance of the data one.

If $m = n$, let $\mathbf{A}$ is an $n \times n$ matrix of rank n, and $\mathbf{A}^{-1}$ exists. Equation 2.27 has the unique solution $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$. However, we do not know $\mathbf{A}$, and therefore $\mathbf{A}^{-1}$ cannot be determined by direct calculation. However there exists a $\mathbf{W}$,

such that $\mathbf{Wx} = \mathbf{WAs} = \mathbf{s}$, and $\mathbf{W}$ happens to be $\mathbf{A}^{-1}$. Now we will embark on an algorithm called FastICA, which starts with an arbitrary $\mathbf{W}$, but rapidly converges to $\mathbf{A}^{-1}$ (Hyvarinen and Oja, 2000).

The first step of the FastICA is whitening. Whitening has two steps. The first step normalizes the signals, making all of them to have mean zero. Take any single channel signal. Determine its mean and subtract it from the value at each time point. The signal thus obtained is of zero mean. The second step of whitening is to make the signals pairwise statistically independent. Statistical independence means that the covariance between any two different signals will be zero. In the mathematical formulation we will have to achieve

$$E\left(\mathbf{x}\mathbf{x}^T\right) = \mathbf{I}, \tag{2.29}$$

where
E denotes expectation
$\mathbf{I}$ denotes the identity matrix

Obviously $E(\mathbf{x}\mathbf{x}^T)$ is a square matrix. Assuming that this square matrix is canonically decomposable (see any standard linear algebra text, such as Hoffman and Kunz, for the definition of canonical decomposition of matrices), we can write

$$E\left(\mathbf{x}\mathbf{x}^T\right) = BDB^T, \tag{2.30}$$

where
B is an orthogonal matrix
D is the eigenvalue matrix of $E(\mathbf{x}\mathbf{x}^T)$

If we replace the initial $\mathbf{x}$ according to the formula (2.31), then the new $\mathbf{x}$ on the left of (2.31) will satisfy (2.29).

$$\mathbf{x} = BD^{-1/2}B^T\mathbf{x}. \tag{2.31}$$

Next, the data will have to be non-Gaussian. A multivariate Gaussian density function is of the form $A\exp\left\{-B(x_1^2 + \cdots + x_n^2)\right\}$. $x_1^2 + \cdots + x_n^2 = r^2$ denotes a spherical shell for real values of r. A spherical shell is uniform in all directions, and therefore there is no preferred direction along which the data can be distinguished from the others. For the same reason, say if only in one variable, say x_1, the data classes are non-Gaussian, they can still be separated. If out of m classes only k ($k < m$) are non-Gaussian and the remaining $m - k$ are Gaussian, then only k classes can be separated by ICA, and the remaining cannot. Filtering the data before running the ICA will be another important step for the success of the ICA algorithm, else the noise may introduce error into the components to be found (noise may obscure a data component, or a noisy component may turn out to be a data component).

To estimate one of the independent components in (2.27) (as after preprocessing (2.28) has been reduced to (2.27)), we consider a linear combination of the x_i's denoted by

$$y = \mathbf{w}^T\mathbf{x} = \sum_i w_i x_i, \tag{2.32}$$

where $\mathbf{w}$ is a vector to be determined. Had $\mathbf{w}$ been one of the rows of $\mathbf{A}$ in (2.27), the linear combination (2.32) would have been one of the independent components that we have been looking for. Since we have no knowledge of $\mathbf{A}$ we cannot determine $\mathbf{w}$, but we can make an estimate of $\mathbf{w}$, and any estimation technique for estimating $\mathbf{w}$ is called an ICA technique. Let us now apply a change of variable $\mathbf{z} = \mathbf{A}^T\mathbf{w}$. Then we have

$$y = \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{A}\mathbf{s} = \mathbf{z}^T\mathbf{s}. \tag{2.33}$$

y is thus a linear combination of s_i with weights given by z_i.

Let us assume that s_i are identically distributed independent (iid) random variables, then the non-Gaussianity of any single random variable will be less than the sum of any number of random variables, which includes this variable (this can be shown easily by calculating the kurtosis under the iid assumption. See Hyvarinen and Oja, 2000, p. 7). In other words, the Gaussianity of y in (2.33) will be the least when it reduces to a component, that is, s_i multiplied by a real number, for some i, or, the non-Gaussianity of y will have to be maximum when it will reduce to a component.

We will measure the non-Gaussianity by negentropy. A Gaussian variable has the largest entropy among all random variables with equal variance (for a proof, see Cover and Thomas, 1991). *Negentropy* $J(\cdot)$ is defined as

$$J(y) = H(y_{gauss}) - H(y), \tag{2.34}$$

where $H(\cdot)$ is entropy. Clearly, y will reduce to a component when $J(y)$ is maximum. FastICA will help to 'rapidly' find a $\mathbf{w}$ (in (2.32)) such that $J(y)$ becomes maximum. FastICA approximates $y = \mathbf{w}^T\mathbf{x}$ such that $J(y)$ reaches close to the maximum. The core steps of FastICA are as follows (Hyvarinen and Oja, 2000):

1. Initialize $\mathbf{w} = \mathbf{w}_0$, where $\mathbf{w}_0$ is an arbitrary nonzero weight vector, with no zero entries.
2. $\mathbf{w}_+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$.
3. $\mathbf{w} = \dfrac{\mathbf{w}_+}{\|\mathbf{w}_+\|}$.
4. If not converged, go back to 2. Convergence means $\mathbf{w} \cdot \mathbf{w}_+ \approx 1$, where $\cdot$ denotes the dot product.

For fast convergence, the function g can be taken in either of the following two ways:

$$\left.\begin{aligned} g(u) &= \tanh(u) \\ g(u) &= u\exp\left(-u^2/2\right) \end{aligned}\right\}$$

Both the choices are somewhat heuristic and cannot be fully justified by rigorous mathematical arguments. Step 2 is the most crucial step of this algorithm. This is the step that approximates (but does not exactly represent) the negentropy as in (2.34).

Before finishing with FastICA let us explain the four steps above. The algorithm tries to converge the weight vector **w** exponentially fast. By step 3 each iteration of **w** is normalized so as to have the unit norm. This means that by virtue of convergence the requirement norm of $\mathbf{w}_+$ will have to be close to 1. In other words, the weight vector can only apply a rotation on the data **x** in order to have y with optimum kurtosis. This qualifies y to become a component (a source component multiplied by a scalar). Geometrically, it is as follows: Consider viewing an egg when its major axis (the axis along the egg's length) is pointing at a normal to your eye. View it from a distance, and take a two dimensional photograph. It will look like a solid white circle. Now apply a rotation to the egg so that its major axis becomes tangential to your eyeball (actually becomes parallel to a tangent to the eyeball). Now you have definitely got an independent component of the data (assuming that the egg is the collection of three-dimensional data points). When the major axis was pointing at a normal to the eyeball, the data set looked spherical and therefore component identification was not certain. But in the second position, existence of a major and a minor component is obvious (respectively, the major and minor axis of the ellipse as a two-dimensional projection of the ellipsoid).

A comparison of Figure 2.9 (raw EEG) with Figure 2.10 (ICA decomposed EEG of Figure 2.9) shows us the efficacy of FastICA as implemented in EEGLAB.

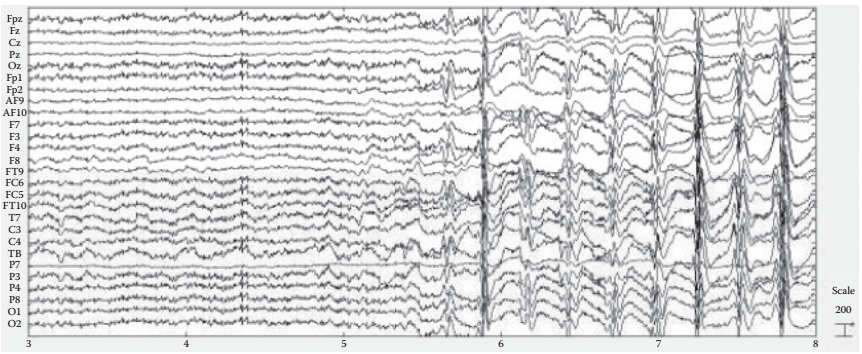


FIGURE 2.9
Snapshot of a 27-channel scalp EEG recording of an epilepsy patient as displayed in EEGLAB.

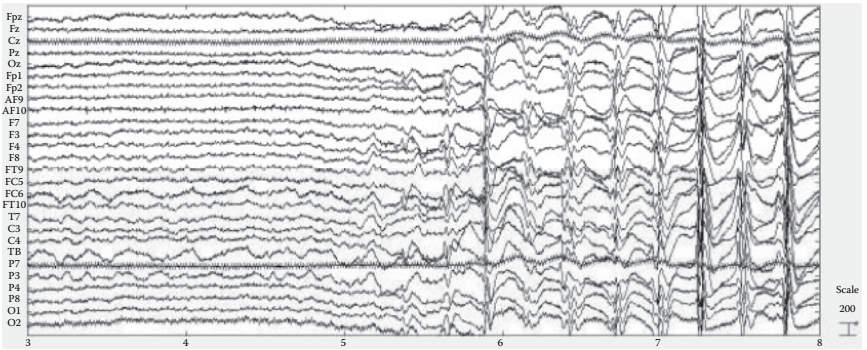


FIGURE 2.10
The EEG signals of Figure 2.8 after being decomposed by FastICA as implemented in EEGLAB.

In Figure 2.9, muscle and 50 Hz line noise artifacts are intermingled with neural signals in such a way that they cannot be easily separated by visual inspection. But after performing the ICA decomposition by FastICA, in Figure 2.10, the muscular artifacts (first two signals (ICA components) from top) and the 50 Hz line noise (the third from the top and the sixth from the bottom signals (ICA components)) are clearly identifiable.

There is another popular way to solve (2.28), called *information maximization* or *InfoMax* (Bell and Sejnowski, 1995). The main quantity to be studied for FastICA was negentropy. For InfoMax it will be mutual information $I(\mathbf{s},\mathbf{x})$ between the source vector $\mathbf{s}$ and the sensor vector $\mathbf{x}$. It is a matter of common sense that the more the mutual information (information or content common in both) in between the two, the better the sources that can be recovered from the sensors. So our goal will be to maximize $I(\mathbf{s},\mathbf{x})$, which is the essence of InfoMax. $I(\mathbf{s},\mathbf{x})$ can be defined in many different ways (Cover and Thomas, 2006), but the most intuitive way to do so is by following the relation between joint entropy $H(\mathbf{s},\mathbf{x})$ and individual entropies $H(\mathbf{s})$ and $H(\mathbf{x})$

$$H(\mathbf{s},\mathbf{x}) = H(\mathbf{s}) + H(\mathbf{x}) - I(\mathbf{s},\mathbf{x}). \tag{2.35}$$

Or

$$\begin{aligned} I(\mathbf{s},\mathbf{x}) &= H(\mathbf{s}) + H(\mathbf{x}) - H(\mathbf{s},\mathbf{x}) \\ &= -\sum P(\mathbf{s}) \log P(\mathbf{s}) - \sum P(\mathbf{x}) \log P(\mathbf{x}) + \sum_s \sum_x P(\mathbf{s},\mathbf{x}) \log P(\mathbf{s},\mathbf{x}) \\ &= -\sum P(\mathbf{s}) \log P(\mathbf{s}) - \sum P(\mathbf{x}) \log P(\mathbf{x}) + \sum_s \sum_x P(\mathbf{s},\mathbf{x}) \log P(\mathbf{s}|\mathbf{x}) P(\mathbf{x}) \\ &= -\sum P(\mathbf{s}) \log P(\mathbf{s}) - \sum P(\mathbf{x}) \log P(\mathbf{x}) + \sum_s \sum_x P(\mathbf{s},\mathbf{x}) \log P(\mathbf{s}|\mathbf{x}) \\ &\quad + \sum_s \sum_x P(\mathbf{s},\mathbf{x}) \log P(\mathbf{x}). \end{aligned}$$

But $\sum_s \sum_x P(\mathbf{s}, \mathbf{x}) \log P(\mathbf{x}) = \sum_x \sum_s P(\mathbf{s}, \mathbf{x}) \log P(\mathbf{x}) = \sum_x P(\mathbf{x}) \log P(\mathbf{x})$ and therefore the above equation becomes

$$I(\mathbf{s},\mathbf{x}) = -\sum P(\mathbf{s}) \log P(\mathbf{s}) + \sum_s \sum_x P(\mathbf{s},\mathbf{x}) \log P(\mathbf{s}|\mathbf{x}) = H(\mathbf{s}) - H(\mathbf{s}|\mathbf{x}). \quad (2.36)$$

This comes from the definition of $H(\mathbf{s}\,|\,\mathbf{x})$ (Cover and Thomas, 2006, p. 17). $H(\mathbf{s}\,|\,\mathbf{x})$ is the *conditional entropy* of $\mathbf{s}$ given $\mathbf{x}$. Here one basic assumption is that both $\mathbf{s}$ and $\mathbf{x}$ have the same probability density P.

The next step will be to maximize $I(\mathbf{s},\mathbf{x})$ as expressed in (2.36), with respect to the noise $\mathbf{n}$ in (2.28) by the following operation

$$\frac{\partial}{\partial \mathbf{n}} I(\mathbf{s},\mathbf{x}) = \frac{\partial}{\partial \mathbf{n}} H(\mathbf{s}), \quad (2.37)$$

because $H(\mathbf{s}\,|\,\mathbf{x})$ does not depend on $\mathbf{n}$. It is clear from (2.28) that $\mathbf{x}$ does not depend on $\mathbf{n}$, and, therefore, none of $P(\mathbf{x})$ and $P(\mathbf{s}\,|\,\mathbf{x})$ depends on $\mathbf{n}$. This implies the joint probability density $P(\mathbf{s},\mathbf{x})$ also does not depend on $\mathbf{n}$. All these together imply that $H(\mathbf{s}\,|\,\mathbf{x})$ (see (2.36)) does not depend on $\mathbf{n}$. Notice that in (2.37) $\partial/\partial\mathbf{n}$ signifies the directional derivative along the vector $\mathbf{n}$ (for directional derivative, see Apostol, 2002). The main advantage of directional derivative with respect to $\mathbf{n}$ is the elegant relation between $I(\mathbf{s},\mathbf{x})$ and $H(\mathbf{s})$ given by (2.37). By symmetry we can also have

$$\frac{\partial}{\partial \mathbf{n}} I(\mathbf{x},\mathbf{s}) = \frac{\partial}{\partial \mathbf{n}} H(\mathbf{x}). \quad (2.38)$$

Let us now consider a more general case than (2.28) of the form

$$\mathbf{x} = \mathbf{g}(\mathbf{W}\mathbf{s} + \mathbf{n}), \quad (2.39)$$

where

$\mathbf{g}$ is a nonlinear function

$\mathbf{W}$ is the *weight matrix* or *mixing matrix*

To be more precise, $\mathbf{g}$ is a vector-valued function of a vector variable (Apostol, 2002), that is, $\mathbf{g} = (g_1, g_2, \ldots, g_n)$, where the dimension of the space of $\mathbf{s}$ = the dimension of the space of $\mathbf{x} = n$. Each component function g_i is a scalar-valued function of a vector variable. In this case, both $I(\mathbf{s},\mathbf{x})$ and $H(\mathbf{s})$ are maximized when we align the high-density part of the probability density function of $\mathbf{s}$ with the highly sloping parts of $\mathbf{g}(\mathbf{s})$ (Bell and Sejnowski, 1995). When all components of $\mathbf{g}$ are highly sloping (as, for example, when each component of $\mathbf{g}$ is a logistic regression function in the corresponding component of $\mathbf{s}$ of the form $g_i = \frac{1}{1+\exp(-s_i)}$), that is, monotonically increasing

or decreasing with respect to each component of **s**, then **g** becomes invertible or there is a one-to-one correspondence between **s** and **x**. Under this one-to-one correspondence, **s** and **x** have the following relationship between their respective probability distribution functions, say $f_{\mathbf{s}}$ and $f_{\mathbf{x}}$.

$$f_{\mathbf{x}} = \frac{f_{\mathbf{s}}}{|J|}, \tag{2.40}$$

where $|J|$ is the absolute value of the Jacobian

$$J = \det\begin{bmatrix} \frac{\partial g_1}{\partial s_1} & \cdots & \frac{\partial g_1}{\partial s_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_n}{\partial s_1} & \cdots & \frac{\partial g_n}{\partial s_n} \end{bmatrix} = (\det \mathbf{W}) \prod_{i=1}^{n} g_i'. \tag{2.41}$$

(Equation (49) of Bell and Sejnowski, 2006). If g_i is a logistic regression in s_i, then

$$g_i' = g_i(1 - s_i). \tag{2.42}$$

Form (2.40) we get

$$H(\mathbf{x}) = -E\left(\log f_{\mathbf{x}}(\mathbf{x})\right) = E\left(\log |J|\right) - E\left(\log f_{\mathbf{s}}(\mathbf{s})\right). \tag{2.43}$$

If $H(\mathbf{x})$ is to be maximized (leading to maximization of mutual information by (2.38)) by adjusting the weights, then only $E(\log|J|)$ in (2.43) is to be considered, because only J is dependent on **W** (by (2.41)). Note that $E(\log|J|) = \int f_{\mathbf{x}}(\mathbf{x}) \log|J| \, d\mathbf{x} = \log|J|$. Similarly, $E(\log(f_{\mathbf{s}}(\mathbf{s})) = \log(f_{\mathbf{s}}(\mathbf{s}))$. Also $H(\mathbf{y})$ is optimized by adjusting **W** only. So

$$\Delta \mathbf{W} \quad \propto \quad \frac{\partial H}{\partial \mathbf{W}} = \frac{\partial}{\partial \mathbf{W}} \log|J| = \frac{\partial}{\partial \mathbf{W}} \log|\det \mathbf{W}| + \frac{\partial}{\partial \mathbf{W}} \log \prod_{i=1}^{n} g_i'. \tag{2.44}$$

The first term in (2.44) is the same, regardless of the transfer function **g**. Since $\det \mathbf{W} = \sum_j cof\ w_{ij}$ for any row i of the weight matrix **W**, for a single weight we have

$$\frac{\partial}{\partial w_{ij}} \log|\det \mathbf{W}| = \frac{cof\ w_{ij}}{\det \mathbf{W}} \quad \text{or} \quad \frac{\partial}{\partial \mathbf{W}} \log|\det \mathbf{W}| = \frac{(adj\ \mathbf{W})^T}{\det \mathbf{W}} = \left[\mathbf{W}^T\right]^{-1}. \tag{2.45}$$

The second term of (2.44) is a sum of logarithmic terms, only one of which depends on a particular w_{ij}, because $g_i = \frac{1}{1 + \exp(-w_{jj} s_i + n_j)}$ when

g_j is a logistic regression (for some other expressions, see (Bell and Sejnowski, 1995)). Clearly,

$$\frac{\partial}{\partial \mathbf{W}} \log \prod_{i=1}^{n} g_i' = (\mathbf{I} - 2\mathbf{g})\mathbf{x}^T. \tag{2.46}$$

Here **I** denotes the vector of 1 with length n. Substituting (2.45) and (2.46) in (2.44) we get

$$\Delta \mathbf{W} \propto \left[\mathbf{W}^T\right]^{-1} + (\mathbf{I} - 2\mathbf{g})\mathbf{x}^T, \tag{2.47}$$

whereas

$$\Delta \mathbf{n} \propto (\mathbf{I} - 2\mathbf{g}). \tag{2.48}$$

Equations 2.47 and 2.48 together constitute the Info-Max algorithm for the separation of blind sources in Equation 2.39, which was used to separate the cortical sources of the scalp EEG (Makeig et al., 1996).

The ICA techniques appear ideally suited for performing source separation in domains where (1) sources are statistically independent (if x, y are random variables, p denotes probability distribution function, and $p(x, y) = p(x) \cdot p(y)$ holds, then x and y are said to be statistically independent), (2) the propagation delays of the 'mixing medium' are negligible, (3) the sources are analogous and have probability density functions that are not too unlike the gradient of a logistic sigmoid, and (4) the number of independent signal sources is the same as the number of sensors (Makeig et al., 1996).

2.2.3 Gist of PCA and ICA Comparison

1. PCA is effective only when components are mutually orthogonal (orthogonality implies independence in the sense of linear algebra). ICA is designed for any independent components, irrespective of whether they are orthogonal or not. In this sense, ICA is a generalization of PCA. In ICA the independence is in the sense of statistical independence. Two linearly independent data sets (vectors) may not be statistically independent, but two statistically independent data sets must be linearly independent. From this point of view also ICA is a generalization of PCA.
2. ICA will NOT work if the data is distributed according to the Gaussian (normal) density function (it cannot be applied at all). But in this case PCA may work (at least it does not have any technical problem with the data being distributed according to a normal density function).
3. PCA will not work well if the eigenvalues of the data covariance matrix are close enough (say, 5, 5.25, 4.93, 5.11, etc.). Here ICA may still work with slightly better results. This happens because as long as the

(independent) components are not equal, the distribution of the (possibly multidimensional) data remains non-Gaussian, which is exploited to decompose the (multidimensional) data into independent components.

4. PCA performs in one go and is therefore relatively more time efficient than ICA. ICA is iterative and therefore takes more time to execute than PCA.
5. PCA and ICA are both matrix operations and therefore can differentiate between 'mixed' (assumed to be a linear mixture of the form $aX+bY+cZ+\cdots$, where X, Y, and Z are independent data components and a, b, and c are weights) components only by rotation and contraction or stretching of the data space. PCA does it blindly, irrespective of the geometric shape of the data space. ICA first identifies where the shape is most convenient to start with and then tries to accentuate this shape on which the component separation is accomplished. For example, if the data set is of ellipsoidal shape, at a distance from some angular view, it will look like a sphere, where all the components will appear equal. But from some other angle the ellipsoidal shape will become very clear and the component along the major axis will be easily identifiable as a dominant component. This 'view' of the data set is the view of maximum (or at least high) non-Gaussianity, which ICA identifies and then decomposes the data set into its (basic) independent components. This decomposition is improved iteratively. On the other hand, PCA never looks for the 'shape' and therefore does not have to bother about non-Gaussianity. It employs classical linear algebraic techniques on the data covariance matrix to find the optimum variance in the multidimensional data set. It works even if there is no variation at all (i.e., all variances are simultaneously optimum).
6. PCA has only one implementation. ICA has multiple implementations (like FastICA, Minimax, etc.).
7. PCA is mathematically precise. ICA algorithms usually have heuristic components.

Some of the drawbacks of ICA have been taken care of in a more recent development called independent vector analysis (IVA) (Lee et al., 2008).

References

Ahlfors, L. V., *Complex Analysis*, McGraw Hill International, Singapore, 1979.

Apostol, T. M., *Calculus, Volume II: Multi-Variable Calculus and Linear Algebra, with Applications to Differential Equations and Probability*, 2nd ed., John Wiley & Sons, Inc., New York, 2002.

Bell, A. J. and T. J. Sejnowski, An information-maximization approach to blind source separation and blind deconvolution, *Neural Comput.*, **7(6)**: 1004–1034, 1995.

Cover, T. M. and J. A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, Hoboken, NJ, 2006.

Delorme, A. and S. Makeig, EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, *J. Neurosci. Methods*, **134**: 9–21, 2004.

Hyvarinen, A. and E. Oja, Independent component analysis: Algorithms and applications, *Neural Netw.*, **13**: 411–430, 2000.

Joliffe, I. T., *Principal Component Analysis*, 2nd ed., Springer, New York, 2002.

Lee, J. H., T. W. Lee, F. A. Jolesz, and S. S. Yoo, Independent vector analysis (IVA): Multivariate approach for fMRI group study, *NeuroImage*, **40(1)**: 86–109, 2008.

Majumdar, K. K. and P. Vardhan, Automatic seizure detection in ECoG by differential operator and windowed variance, *IEEE Trans. Neural Syst. Rehabil. Eng.*, **19(4)**: 356–365, 2011.

Makeig, S., A. J. Bell, T.-P. Jung, and T. J. Sejnowski, Independent component analysis of electroencephalographic data, in: D. Touretzky, M. Mozer, and M. Hasselmo (eds.), *Advances in Neural Information Processing Systems 8*, pp. 145–151, MIT Press, Cambridge, MA, 1996.

Parks, T. W. and J. H. McClelland, Chebyshev approximation for nonrecursive digital filters with linear phase, *IEEE Trans. Circuit Theory*, **19(2)**: 189–194, 1972.

Proakis, J. G. and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, 4th ed., Pearson Education Inc., New Delhi, India, 2007.

Raja, J., B. Muralikrishnan, and S. Fu, Recent advances in separation of roughness, waviness and form, *Precession Eng.*, **26**: 222–235, 2002.

Rao, D. G. and V. P. Gejji, *Digital Signal Processing: Theory and Lab Practice*, 2nd ed., Pearson Education Inc., New Delhi, India, 2010.

Shlens, J., A tutorial on principal component analysis, 2009, available at http://www.snl.salk.edu/~shlens/pca.pdf (accessed April 2, 2015).

Smith, L. I., A tutorial on principal component analysis, 2002, available at http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf (accessed April 2, 2015).

3

Source Localization

Localization of cortical sources in scalp EEG is of fundamental importance in EEG research. Our brain functions vary within milliseconds, which is electrically transmitted to the scalp electrodes almost instantaneously and recorded in the form of EEG (see Chapter 1 for details). Therefore, EEG has excellent temporal resolution. On the other hand, the brain is divided into functionally specialized areas (Brodmann's areas and their refinements) and therefore knowledge of the exact source of the signal is important for understanding the EEG signal's task-specific significance.

The EEG source localization problem has two major components: forward problem and inverse problem. In the forward problem, a model of the head is to be built up. Once this is done with the details of electrical conductivity of the head tissues, the forward problem is resolved. In the inverse problem, the recorded scalp potential in individual EEG channels is to be studied to find the location of the cortical electrical activities responsible for the electrical signals in the scalp EEG channels.

3.1 Forward Problem

The purpose of the forward problem is to make a model of the head of the subject whose EEG is being collected. This model consists in studying the electrical impedance of the head tissues (see Section 1.2), so that when an electrical source is given in the cortex, one can calculate how its electrical field propagates to the scalp. There are at least nine clearly identifiable tissue layers between the cortical surface and the scalp electrodes (Figure 1.8). Each tissue layer is spatially nonhomogeneous and therefore has different electrical impedance at different parts.

3.1.1 Boundary Element Method

Although in reality impedance over an entire tissue layer is nonhomogeneous, if we ignore this fact during modeling and computation and treat the entire tissue layer to have a uniform impedance value, we can model the head by boundary element method (BEM). In most head models, three tissue layers

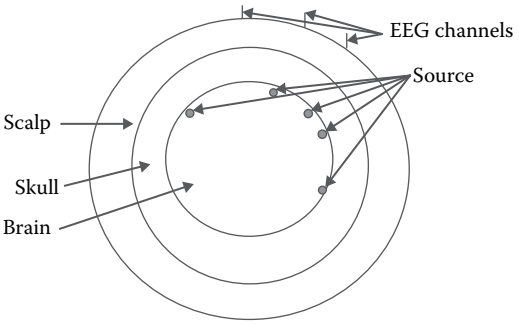


FIGURE 3.1
Schematic diagram of spherical head model with three tissue layers—brain, skull, and scalp. Also shown are electric sources in the brain (cortex) and EEG channels on the scalp. The thickness of the tissues is not to scale.

are considered most important: brain (cortex), skull, and scalp. The other tissues are ignored. For a schematic diagram of this model, see Figure 3.1.

From Figure 3.1, it is evident that the electric field has to propagate from the electric sources in the brain through the brain, skull, and scalp to the EEG electrodes. This field is an alternating current field with irregular frequencies. Knowing the impedance of the tissues is imperative for modeling the propagation of the field from the source to the EEG electrodes. Impedance of human head tissues was traditionally measured in cadavers, but this is not the same as the impedance in a live brain. It is also possible to measure impedances in a live brain by stimulating the tissues noninvasively with small amounts of electric current pulses. Among all the tissue layers in the head, the skull has the highest impedance. The skull also has a variable thickness and it is more porous in some places than others. In reality, it has a highly inhomogeneous impedance. Nevertheless, in a BEM model, its impedance is assumed to be uniform everywhere.

3.1.1.1 Dipole Source Model

In Figure 3.2, Ω_{N+1} is the air and its conductivity σ_{N+1} is zero. The electric potential on the scalp is generated due to electric 'dipole' source in the brain. In Chapter 1, we have seen pyramidal neurons are oriented normal to the cortical surface (Figure 1.3). When a number of them within a small neighborhood fire synchronously, they form an electric dipole, as shown in Figure 3.3a (compare Figure 1.3 with Figure 3.3a). When this number reaches the order of tens of thousands, the potential at the scalp reaches the order of tens of microvolts.

The dipole vector $\mathbf{d}$ is represented as

$$\mathbf{d} = d_x\mathbf{e}_x + d_y\mathbf{e}_y + d_z\mathbf{e}_z, \tag{3.1}$$

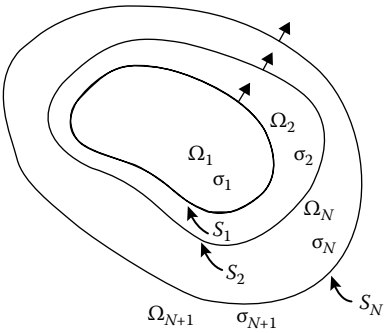


FIGURE 3.2
A schematic diagram of a realistic (not spherical) head model, where surface S_1 is the cortical surface and surface S_N is the scalp. Ω_i and σ_i are, respectively, the bulk or volume of the tissue and the conductivity (assumed to be uniform) of the tissue enclosed between S_{i-1} and S_i for $i \geq 2$. Arrows indicate direction of electric field propagation. (Taken from Kybic, J. et al., *Phys. Med. Biol.*, 51, 1333, 2006.)

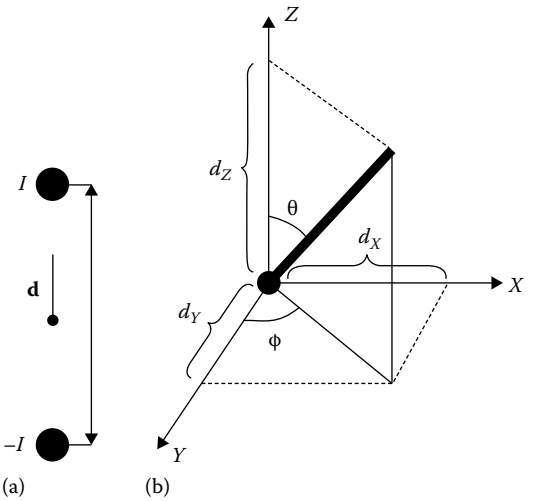


FIGURE 3.3
Schematic representation of an electrical dipole: (a) an electrical dipole vector **d** and (b) *X*, *Y*, and *Z* components of a dipole vector **d**. (Taken from Hallez, H. et al., *J. NeuroEng. Rehab.*, 4, 46, 2007, available at http://www.jneuroengrehab.com/content/4/1/46.)

where $\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$ are unit vectors in the *X*, *Y*, and *Z* directions, respectively (Figure 3.3b). A dipole needs six parameters to be fully specified, three for location in the three-dimensional space and d_x, d_y, d_z for orientation, where $\|\mathbf{d}\| = \sqrt{d_x^2 + d_y^2 + d_z^2}$. Potential $V(\mathbf{r}_i, \mathbf{r}_{dip_j}, d_j\mathbf{e}_{d_j})$ at the *i*th electrode positioned at $\mathbf{r}_i = \mathbf{e}_x x_i + \mathbf{e}_y y_i + \mathbf{e}_z z_i$ on the scalp due to the *j*th dipole $d_j\mathbf{e}_{d_j}$ at the position $\mathbf{r}_{dip_j}$ in the brain can be decomposed as

$$V\left(\mathbf{r}_i, \mathbf{r}_{dip_j}, d_j\mathbf{e}_{d_j}\right) = d_x V\left(\mathbf{r}_i, \mathbf{r}_{dip_j}, \mathbf{e}_{d_j}\right) + d_y V\left(\mathbf{r}_i, \mathbf{r}_{dip_j}, \mathbf{e}_{d_j}\right) + d_z V\left(\mathbf{r}_i, \mathbf{r}_{dip_j}, \mathbf{e}_{d_j}\right). \quad (3.2)$$

TABLE 3.1

The Values of Conductivity σ of Different Tissues in the Human Head Collected from Published Literature

Tissue	Geddes and Baker (1967)	Oostendorp et al. (2000)	Goncalves et al. (2003)	Gutirrez et al. (2004)	Lai et al. (2005)
Scalp	0.43	0.22	0.33	0.749	0.33
Skull	0.006–0.015	0.015	0.0081	0.012	0.0132
Cerebrospinal fluid				1.79	
Brain	0.12–0.48	0.22	0.33	0.313	0.33

Source: Adapted from Hallez, H. et al., *J. NeuroEng. Rehab.*, 4, 46, 2007, available at http://www.jneuroengrehab.com/content/4/1/46.

The current density due to a source is a vector field and can be represented by $\mathbf{J}(x,y,z)$. The unit of the current density is A/m^2. The divergence of the vector field $\mathbf{J}$ is the conduction of current given by

$$\nabla \cdot \mathbf{J} = \lim_{\Omega \to 0} \frac{1}{\Omega} \oint_{\partial\Omega} \mathbf{J} d\mathbf{S} = I = \nabla \cdot (\sigma \nabla V), \tag{3.3}$$

with the boundary condition

$$\sigma \nabla V \cdot \mathbf{e}_n = 0. \tag{3.4}$$

Here Ω is the volume within which the current density is $\mathbf{J}$; $d\mathbf{S} = \mathbf{e}dS$, where dS is the total surface of the infinitesimal volume $d\Omega$ and $\mathbf{e}$ is the unit vector normal to the surface dS; $\mathbf{e}_n$ is normal to the surface between scalp and air; I is the current flowing inwards (known as a current sink and assigned a negative sign) into $\partial\Omega$ or outward (known as current source and assigned a positive sign) from it; V is the potential at the surface dS; σ is the conductivity. Equation 3.3 is known as Poisson's equation. Equation 3.4 holds when dS is the scalp surface, because no current can be injected into the air, whose conductivity is zero.

A set of values for conductivity σ of different head tissues, as reported in the literature, are given in Table 3.1. σ has wide variation across individuals. An incorrect value of σ for any head tissue will become a major source of localization error for cortical current sources (Acar and Makeig, 2013).

Propagation of V through tissue layers, under the assumption that σ remains constant throughout the layer, is governed by the Poisson's equation (3.3). Let us consider the head model in Figure 3.2 in a Cartesian coordinate system. It is clear from Figure 3.3 that an electric source dipole in brain must have a

positive and a negative pole. Let the location of those poles be at (x_2, y_2, z_2) and (x_1, y_1, z_1), respectively. Then (3.3) takes the form

$$\frac{\partial}{\partial x}\left(\sigma\frac{\partial V}{\partial x}\right)+\frac{\partial}{\partial y}\left(\sigma\frac{\partial V}{\partial y}\right)+\frac{\partial}{\partial z}\left(\sigma\frac{\partial V}{\partial z}\right)$$
$$=-I\delta(x-x_1)\delta(y-y_1)\delta(z-z_1)+I\delta(x-x_2)\delta(y-y_2)\delta(z-z_2) \quad (3.5)$$

where δ is Dirac delta function. Similarly, for polar coordinate it becomes

$$\nabla\left(\sigma\nabla(V)\right)=-I\delta(\mathbf{r}-\mathbf{r}_1)+I\delta(\mathbf{r}-\mathbf{r}_2), \quad (3.6)$$

where $\mathbf{r}$ is the radius vector of (x, y, z), etc. According to the underlying assumption of BEM, σ changes at the tissue layer boundary only (Figure 3.2). This change can be modeled in different ways. Here we describe only an elementary model. However, before that let us discuss how this solution is going to contribute to the forward problem. The forward problem is about how the different electrical sources in the brain, when active simultaneously, will generate potentials at all the electrodes on the scalp. Let there be p number of electrical dipole sources in the brain and N number of EEG electrodes on the scalp. Then the forward problem can be formulated in matrix form as

$$\mathbf{V}=\begin{bmatrix}V(\mathbf{r}_1)\\ \vdots \\ V(\mathbf{r}_N)\end{bmatrix}=\begin{bmatrix}g(\mathbf{r}_1,\mathbf{r}_{dip_1},\mathbf{e}_{d_1}) & \cdots & g(\mathbf{r}_1,\mathbf{r}_{dip_p},\mathbf{e}_{d_p})\\ \vdots & \vdots & \vdots \\ g(\mathbf{r}_N,\mathbf{r}_{dip_1},\mathbf{e}_{d_1}) & \cdots & g(\mathbf{r}_N,\mathbf{r}_{dip_p},\mathbf{e}_{d_p})\end{bmatrix}\begin{bmatrix}d_1\\ \vdots \\ d_p\end{bmatrix}=G(\mathbf{r}_j,\mathbf{r}_{dip_i},\mathbf{e}_{d_i})\begin{bmatrix}d_1\\ \vdots \\ d_p\end{bmatrix}, \quad (3.7)$$

where $g(\mathbf{r}_i, \mathbf{r}_{dip_j}, \mathbf{e}_{d_j})$ is the potential at the point $\mathbf{r}_i=\mathbf{e}_x x_i+\mathbf{e}_y y_i+\mathbf{e}_z z_i$ or (x_i, y_i, z_i) on the scalp due to the jth dipole at the position $\mathbf{r}_{dip_j}$ with orientation along the unit vector $\mathbf{e}_{d_p}$ and magnitude d_p (see (3.2)). Obviously, the potential $V(\mathbf{r}_i)$ at the electrode position $\mathbf{r}_i$ on the scalp is due to linear superposition (linear addition) of effects due to all the p dipoles being active in a given time instance. The matrix

$$G(\mathbf{r}_j,\mathbf{r}_{dip_i},\mathbf{e}_{d_i})=\begin{bmatrix}g(\mathbf{r}_1,\mathbf{r}_{dip_1},\mathbf{e}_{d_1}) & \cdots & g(\mathbf{r}_1,\mathbf{r}_{dip_p},\mathbf{e}_{d_p})\\ \vdots & \vdots & \vdots \\ g(\mathbf{r}_N,\mathbf{r}_{dip_1},\mathbf{e}_{d_1}) & \cdots & g(\mathbf{r}_N,\mathbf{r}_{dip_p},\mathbf{e}_{d_p})\end{bmatrix} \quad (3.8)$$

is called *gain matrix*. The entire purpose of the forward problem is to calculate the gain matrix (3.8). Then in the inverse problem, scalp potential $V(\mathbf{r}_i)$ at the electrode placed in position $\mathbf{r}_i$ for all i is known and (3.7) is solved to get $d_1, \ldots, d_p$.

For N electrodes, p dipoles, and T discrete time points, $\mathbf{V}$ in (3.7) becomes a matrix:

$$\mathbf{V} = \begin{bmatrix} V(\mathbf{r}_1,1) & \ldots & V(\mathbf{r}_1,T) \\ \vdots & \vdots & \vdots \\ V(\mathbf{r}_N,1) & \ldots & V(\mathbf{r}_N,T) \end{bmatrix}.$$

Then (3.7) takes the form

$$\mathbf{V} = G\left(\mathbf{r}_i, \mathbf{r}_{dip_j}, \mathbf{e}_{d_j}\right) \begin{bmatrix} d_{1,1} & \ldots & d_{1,T} \\ \vdots & \vdots & \vdots \\ d_{p,1} & \ldots & d_{p,T} \end{bmatrix} = G\left(\mathbf{r}_i, \mathbf{r}_{dip_j}, \mathbf{e}_{d_j}\right)\mathbf{D}, \tag{3.9}$$

where $\mathbf{D} = \begin{bmatrix} d_{1,1} & \ldots & d_{1,T} \\ \vdots & \vdots & \vdots \\ d_{p,1} & \ldots & d_{p,T} \end{bmatrix}$. Equation 3.9 is the formulation of the most general form of the forward problem, the *time-varying forward problem*.

The Poisson equation (3.3) in its variant (3.5) (for Cartesian coordinate system) or (3.6) (for polar coordinate system) is solved to get $g(\mathbf{r}_i, \mathbf{r}_{dip_j}, \mathbf{e}_{d_j})$ for all i, j. For a numerical solution we need to subdivide the tissue boundary surfaces into triangular grids (Figure 3.4). The surface S_j is triangulated into N_j number of triangles (Figure 3.4). At any point $\mathbf{r}$ on that surface the potential is calculated using (3.10). Note that in most head models $R = 3$ and conductivity of air is zero.

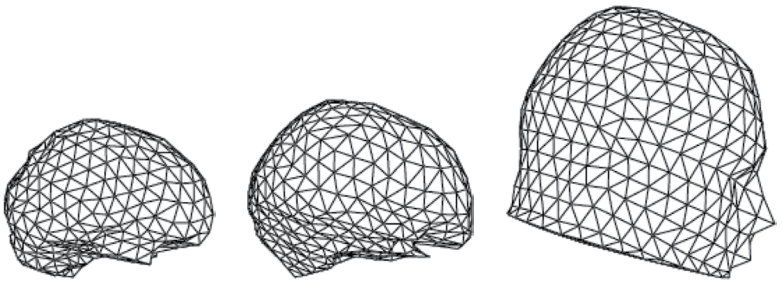


FIGURE 3.4
Triangulation of human head tissue surfaces. Leftmost is the brain, middle is the skull, and rightmost is the scalp. (Taken from Hallez, H. et al., *J. NeuroEng. Rehab.*, 4, 46, 2007, available at http://www.jneuroengrehab.com/content/4/1/46.)

$$V(\mathbf{r}) = \frac{2\sigma_1}{\sigma_j + \sigma_{j+1}} V_1(\mathbf{r}) + \frac{1}{2\pi} \sum_{k=1}^{R} \frac{\sigma_k - \sigma_{k+1}}{\sigma_j + \sigma_{j+1}} \sum_{l=1}^{N_j} \int_{\mathbf{r}' \in \Delta \mathbf{S}_{l,j}} V(\mathbf{r}') \frac{\mathbf{r}' - \mathbf{r}}{\|\mathbf{r}' - \mathbf{r}\|^3} \cdot d\mathbf{S}_l, \quad (3.10)$$

where

σ_1 is the conductivity of the medium in which the dipole is located (brain)

$V_1(\mathbf{r})$ is the potential at $\mathbf{r}$ for an infinite medium with conductivity σ_1 (Figure 3.2)

$\Delta \mathbf{S}_{l,j}$ is the lth triangle of the jth surface

$d\mathbf{S}_l$ is a vector normal to the surface element dS_l (arbitrarily small) within the lth triangle of the surface S_j, whose value is equal to the area of dS_l

$\cdot$ is the dot product

The potential calculated in each elementary triangle $\Delta \mathbf{S}_{l,j}$ is given by

$$\int_{\mathbf{r}' \in \Delta \mathbf{S}_{l,j}} V(\mathbf{r}') \frac{\mathbf{r}' - \mathbf{r}}{\|\mathbf{r}' - \mathbf{r}\|^3} \cdot d\mathbf{S}_l.$$

This potential is assumed to act on the center of mass of the elementary triangle.

Equation 3.10 has to be solved numerically, and $V(\mathbf{r}_i)$ on each triangle on the surface between scalp and air will give $\sum_j g(\mathbf{r}_i, \mathbf{r}_{dipj}, \mathbf{e}_{dj}) d_j$.

3.1.1.2 Distributed Source Model

Despite being an elegant model, the main disadvantage of the dipole source model is that the exact number of dipoles will have to be known in advance. Active cortical sources are time varying. They periodically become hyperactive and almost inactive in the order of tens of milliseconds. Determining the exact number of sources within a time window is extremely difficult. Often locations of dominant sources are estimated depending on the task, which is likely to involve brain areas at a certain time lag. An alternative is to assume electrical sources at all the grid points in the cortical mesh (the leftmost mesh in Figure 3.4) or the surface of the cortex. These sources are modeled as monopole (either has positive or negative charge) normal to the surface. This is called *distributed source model*. Obviously there are as many sources as there are points in the mesh (vertices of the tessellated triangles). One disadvantage is the number of sources becomes very high and the size of the gain matrix becomes large, prolonging the time to compute the gain matrix. The forward problem (3.7) becomes

$$\mathbf{V} = \begin{bmatrix} V(\mathbf{r}_1) \\ \vdots \\ V(\mathbf{r}_N) \end{bmatrix} = \begin{bmatrix} g(\mathbf{r}_1, \mathbf{r}_{source_1}) & \cdots & g(\mathbf{r}_1, \mathbf{r}_{source_p}) \\ \vdots & \vdots & \vdots \\ g(\mathbf{r}_N, \mathbf{r}_{source_1}) & \cdots & g(\mathbf{r}_N, \mathbf{r}_{source_p}) \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_p \end{bmatrix} = G(\mathbf{r}_j, \mathbf{r}_{source_i}) \begin{bmatrix} d_1 \\ \vdots \\ d_p \end{bmatrix}, \quad (3.11)$$

where
p is the number of sources
$\mathbf{r}_{source_j}$ is the position of the jth source

Notice that the source dipoles are all normal to the surface S_1 (assumed to be smooth) and therefore the pole away from the surface does not exert any effect on the scalp electrodes. Each source works as an effective monopole, all of which have the same orientation and therefore orientation does not need to be considered as a variable. $g(\mathbf{r}_i, \mathbf{r}_{source_j})$ can be determined in different ways. In this case

$$v_{S_{k,j}}(\mathbf{r}) = \frac{1}{4\pi} \int_{\mathbf{r}' \in S_{k,j}} \frac{\mathbf{n}_{S_{k,j}}(\mathbf{r}') \cdot (\mathbf{r} - \mathbf{r}')}{\|\mathbf{r} - \mathbf{r}'\|^3} J_{S_{k,j}}(\mathbf{r}') d\mathbf{r}', \quad (3.12)$$

where
$v_{S_{k,j}}(\mathbf{r})$ is the potential on the triangular patch $S_{k,j}$ of the cortical surface
$\mathbf{n}_{S_{k,j}}$ is a vector normal to elementary surface $S_{k,j}$
$J_{S_{k,j}}(\mathbf{r}')$ is the current density on the elementary surface $S_{k,j}$ (Kybic et al. 2005, equation (11))

Substituting (3.12) in (3.10), we get

$$V(\mathbf{r}) = \frac{2\sigma_1}{\sigma_j + \sigma_{j+1}} V_1(\mathbf{r}) + \frac{1}{4\pi} \sum_{k=1}^{R} \frac{\sigma_k - \sigma_{k+1}}{\sigma_j + \sigma_{j+1}} \sum_{l=1}^{N_j} \int_{\mathbf{r}' \in S_{k,j}} \frac{\mathbf{n}_{S_{k,j}}(\mathbf{r}') \cdot (\mathbf{r} - \mathbf{r}')}{\|\mathbf{r} - \mathbf{r}'\|^3} J_{S_{k,j}}(\mathbf{r}') d\mathbf{r}'. \quad (3.13)$$

It may be informative to note that there is an open source software called OpenMEEG available for solving the forward problem by BEM (OpenMEEG 2.1, 2011). Some toy data sets of three-layer spherical head models have also been provided with the software. It is advisable to generate head models under different constraints using the software and the data sets in order to have a firsthand feel of how some BEM solutions to the forward problem work. Numerous free and commercial software are available for solving the source localization problem as a whole.

3.1.2 Finite Element Method

In the last section, we considered values of tissue conductivity only at the tissue boundaries and assumed that these values remain uniform throughout the layer of one particular tissue type. But this is simplistic. For example, the skull has different thickness and different porosity at different places. A more realistic head model will be the one in which the skull is subdivided into smaller pieces, each having a different conductivity. If the pieces are small enough, the conductivity may reasonably be taken to be uniform across that small piece. This is more or less true for other tissues as well. The method that is invoked to solve (3.3) on a tessellation of the entire three-dimensional tissue volume is called *finite element method* (FEM) (Figure 3.5). Here we will describe FEM for distributed sources. The modification of the method for a given number of dipole sources at specific locations will be straightforward.

Note that (3.3) and (3.4) together is the fundamental form of forward problem of the brain electrical source imaging. The relation between σ and V is a complicated nonlinear mapping. However, the first-order variation in the scalar potential function ΔV due to perturbation $\Delta\sigma$ in the conductivity distribution can be determined. Let V and V_0 be the potential functions for the

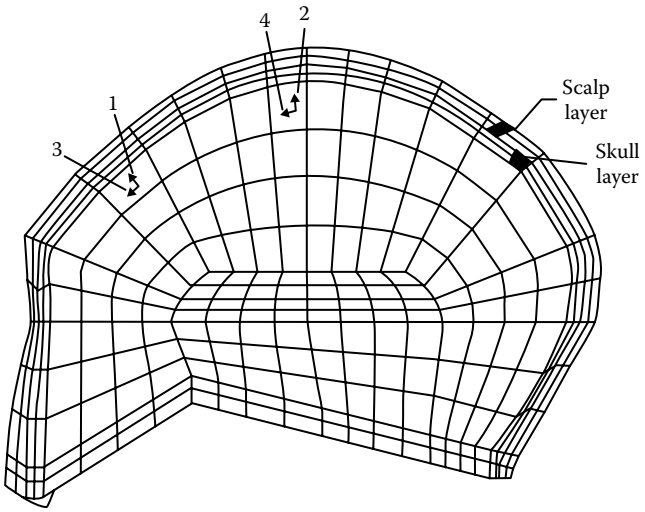


FIGURE 3.5
FEM model of a mid-sagittal section of a human head. 1, 2, 3, and 4 are dipole source locations. Each square is an element with fixed conductivity and each vertex is a node having an electric potential due to the electric source. The skull and scalp both have two layers: one inner and one outer. (Taken from Yan, Y. et al., *Med. Biol. Eng. Comput.*, 29, 475, 1991.)

conductivity σ and σ_0, respectively. Then, they satisfy the following differential equations (Gencer and Acar, 2004):

$$\nabla\left(\sigma_1 \nabla V_1\right) = \nabla \cdot \mathbf{J} \tag{3.14}$$

$$\nabla\left(\sigma \nabla V\right) = \nabla \cdot \mathbf{J}. \tag{3.15}$$

Substituting $V_1 + \Delta V$ and $\sigma_1 + \Delta\sigma$ for V and σ, respectively, in (3.15), we get

$$\nabla \cdot \left[\sigma_1 \nabla V_1 + \sigma_1 \nabla\left(\Delta V\right) + \Delta\sigma \nabla V_1 + \Delta\sigma \nabla\left(\Delta V\right)\right] = \nabla \cdot \mathbf{J}. \tag{3.16}$$

From (3.14) and (3.16), we get

$$\left.\begin{aligned} \nabla \cdot \left(\sigma_1 \nabla\left(\Delta V\right)\right) &= -\nabla \cdot \left(\Delta\sigma \nabla V_1\right) \\ \sigma \nabla\left(\Delta V\right) \cdot \mathbf{e}_n &= 0 \end{aligned}\right\}. \tag{3.17}$$

The second equality in (3.17) is the boundary condition (see also (3.4)). Equation 3.17 implies that the strength of ΔV is determined by the relative change in the tissue conductivity and the electric field calculated for the initially assumed conductivity distribution. Thus, for perturbations at different tissues, as long as the same relative change in conductivity is considered, the strength of the initial electric field becomes the major factor in determining the strength of ΔV.

Assume now that the initial conductivity distribution σ_1 is constant, then we obtain

$$\left.\begin{aligned} \sigma_1 \nabla^2\left(\Delta V\right) &= -\nabla\sigma \cdot \nabla V_1 - \frac{\Delta\sigma}{\sigma_1}\left(\nabla \cdot \mathbf{J}\right) \\ \sigma \nabla\left(\Delta V\right) \cdot \mathbf{e}_n &= 0 \end{aligned}\right\}. \tag{3.18}$$

Thus, the regions where σ varies due to a perturbation become regions of charge accumulation. The dot product of the electric field due to primary sources and the conductivity gradient determines the strength of these charges. For points in the active area, if there is a conductivity perturbation, then the primary source $\nabla \cdot \mathbf{J}$ weighted by the relative change in conductivity also appears as a source for ΔV.

If one further assumes a uniform perturbation in an initially uniform conducting body, the first term on the right-hand side of (3.18) vanishes (as the gradient of conductivity is equal to zero). Therefore

$$\left.\begin{aligned} \sigma_1 \nabla^2\left(\Delta V\right) &= -\frac{\Delta\sigma}{\sigma_1}\left(\nabla \cdot \mathbf{J}\right) \\ \sigma \nabla\left(\Delta V\right) \cdot \mathbf{e}_n &= 0 \end{aligned}\right\}. \tag{3.19}$$

This equation has the same form of Equations 3.3 and 3.4 taken together, when written for an initially uniform conductivity distribution σ_1. The only difference is the minus sign on the right-hand side and the scaling factor

$\Delta\sigma/\sigma_1$ before the source term. Thus, any relative increase in conductivity decreases the potential by the same amount and vice versa, preserving the trend of the initial potential distribution V_1.

Before concluding this section, we must relate the change in potentials ΔV to the change in the actual measurement. The actual measurements are the potential differences measured between electrodes placed on the scalp surface. This can be obtained by calculating the line integral of the electric field $-\Delta V$ on any line connecting the electrode pairs. Thus, a change in this measurement due to a conductivity perturbation $\Delta\sigma$ is the line integral of $-\nabla(\Delta V)$ between these points.

To solve the forward problem for arbitrary geometries and conductivity distributions, numerical methods are required. Here, Equations 3.3 and 3.4 are solved using FEM. In the FEM, the entire head is discretized into volume elements (a two-dimensional analog of what is shown in Figure 3.5) on which the potential function is defined by either linear or quadratic interpolation functions (Gencer and Acar, 2004). A linear system of equations is obtained either by application of a variational principle or by the technique of weighted residuals. The elements can have various shapes (tetrahedral or hexahedral) allowing the use of an irregular grid. We will follow the FEM formulation with linear and quadratic hexahedral isoparametric elements. This formulation defines the source as a constant volume current density inside the elements (distributed source model). Dipole source model has also been implemented. An equivalent discretized model is constructed for each element using the Galerkin weighted residual method, and then all elements' contributions are assembled to form a large system of equations. For a discretization into m nodes and n elements (the whole head is tessellated into n elements, $m > n$, see Figure 3.5), the system of equations is of the form

$$\mathbf{A}(\sigma)\mathbf{V} = \mathbf{J}, \tag{3.20}$$

where

- $\mathbf{V}$ is an m-dimensional vector of potential at nodes (unknown in forward problem)
- $\mathbf{J}$ is an m-dimensional vector of current density of current sources (see Figure 3.5)
- $\mathbf{A}(\sigma) = \left\{a_{ij}(\sigma)\right\}_{i,j=1}^{m}$ is a symmetric $m \times m$ matrix containing element geometry and conductivity information

$a_{ij}(\sigma)$ is actually dependent on conductivity of all elements and therefore σ is an n-dimensional vector containing conductivity values of all n elements. Note that due to the current sources, each (small) element has a current density and an electric potential at each of the nodes. In an FEM forward problem, we need to know $\mathbf{V}$, while $\mathbf{A}$ and $\mathbf{J}$ are known. Once we know $\mathbf{V}$, we know the potential at each node of an element and therefore the potential at any point of the element as a convex combination of node potentials

in that element. With respect to changing conductivity from one element to another potential change is approximated by

$$\mathbf{V}(\sigma_1 + \Delta\sigma) = \mathbf{V}(\sigma_1) + \mathbf{S}_\mathbf{V}\Delta\sigma, \tag{3.21}$$

$\mathbf{S}_\mathbf{V}$ is the *sensitivity matrix*. Now we will see how the sensitivity matrix is calculated following Gencer and Acar (2004).

The first-order variation in the scalar potential function ΔV related to a perturbation $\Delta\sigma$ in the conductivity distribution was given in (3.17). This equation can be written by substituting $\sigma - \sigma_1$ for $\Delta\sigma$ as follows:

$$\left.\begin{aligned} \nabla\cdot\left(\sigma_1\nabla(\Delta V)\right) &= -\nabla\cdot\left(\sigma_1\nabla V_1\right) + \nabla\cdot\left(\sigma\nabla V_1\right) \\ \sigma\nabla(\Delta V)\cdot\mathbf{e}_n &= 0 \end{aligned}\right\}, \tag{3.22}$$

which is more suitable for numerical implementation. Invoking the matrix notation used in Equation 3.17, Equation 3.22 can be discretized as

$$\mathbf{A}(\sigma_1)\Delta\mathbf{V} = -\mathbf{A}(\sigma)\mathbf{V}_1 + \mathbf{A}(\sigma_1)\mathbf{V}_1. \tag{3.23}$$

If the right-hand side of this equation is expressed as a linear approximation around an initial conductivity σ_0, then the following equations can be obtained:

$$\mathbf{A}(\sigma_1)\Delta\mathbf{V} = -\frac{\partial\mathbf{A}(\sigma)}{\partial\sigma}\mathbf{V}_1\bigg|_{\sigma=\sigma_1}\Delta\sigma. \tag{3.24}$$

Note that $\Delta\sigma = \sigma - \sigma_1$. From (3.24), we get

$$\Delta\mathbf{V} = -\left\{\mathbf{A}(\sigma_1)\right\}^{-1}\frac{\partial\mathbf{A}(\sigma)}{\partial\sigma}\mathbf{V}_1\bigg|_{\sigma=\sigma_1}\Delta\sigma. \tag{3.25}$$

Equation 3.25 defines a linear relation between $\Delta\mathbf{V}$ and $\Delta\sigma$. However, it requires the inverse of $\mathbf{A}$ which is computation intensive (the fastest algorithm known till 2011 takes time $O(m^{2\cdot3727})$ for inverting a $m\times m$ matrix, the naïve algorithm executes in $O(m^3)$ time). The computational requirements can be reduced because only the potential at a subset of the nodes at the electrode positions $\mathbf{V}_s$ is of interest. Let $\mathbf{L}$ be a $l\times m$ sparse matrix that selects l positions to place electrodes out of m nodes (each node is a possible position for an electrode placement) in the FEM mesh, that is, $\mathbf{V}_\mathbf{S} = \mathbf{L}\mathbf{V}$. Then changes in the potentials at the electrode locations $\Delta\mathbf{V}_\mathbf{S}$ can be expressed as

$$\Delta\mathbf{V}_\mathbf{S} = -\mathbf{L}\left\{\mathbf{A}(\sigma_1)\right\}^{-1}\frac{\partial\mathbf{A}(\sigma)}{\partial\sigma}\mathbf{V}_1\bigg|_{\sigma=\sigma_1}\Delta\sigma. \tag{3.26}$$

$$\Delta \mathbf{V_S} = \mathbf{S_V}. \tag{3.27}$$

The right-hand side of Equation 3.26 gives us the computable form of the sensitivity matrix $\mathbf{S_V}$ and (3.27) tells us that $\mathbf{S_V}$ is equal to the change in potential at the scalp electrodes. In the calculation of the sensitivity matrix, since $\mathbf{A}^{-1}$ is symmetric, the term $\mathbf{LA}^{-1}$ can be computed by solving $\mathbf{Ax} = \mathbf{I}_i$ for each column $\mathbf{I}_i$ of $\mathbf{L}$.

3.1.3 Finite Difference Method

There are two different cases for the finite difference method (FDM): (1) isotropic media FDM, or iFDM, and (2) anisotropic media FDM, or aFDM. In iFDM, within a box of one tissue layer (Figure 3.6), conductivity is taken to be homogeneous in all directions like in BEM, and in a FDM tissue, conductivity

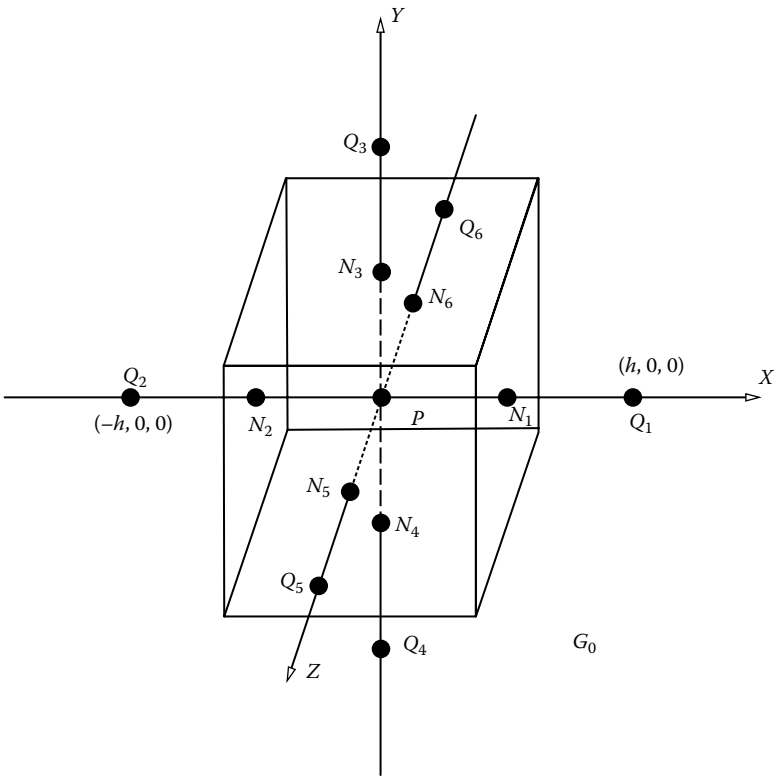


FIGURE 3.6
G_0 is a typical grid in an iFDM computation. P is a node with six neighboring nodes Q_i, $i \in \{1, \ldots, 6\}$. (Taken from Hallez, H. et al., *J. NeuroEng. Rehab.*, 4, 46, 2007, available at http://www.jneuroengrehab.com/content/4/1/46.)

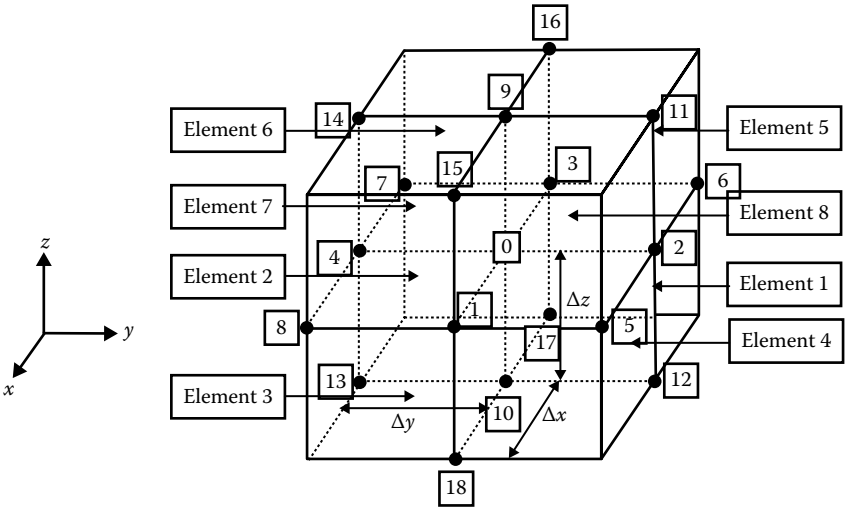


FIGURE 3.7
A typical environment around a node 0. Eighteen neighboring points have been considered as vertices of eight neighboring elements. (Taken from Saleheen, H.I. and Ng, K.T., *IEEE Trans. Biomed. Eng.*, 44(9), 800, 1997.)

is taken to be heterogeneous in different directions within the box (Figure 3.7) like in FEM. The latter is more realistic in a tissue like the skull.

3.1.3.1 iFDM

Consider a typical node P in a cubic grid with internode spacing h. The six neighboring nodes are Q_i, $i \in \{1,\ldots,6\}$ (Figure 3.6). Now let us introduce α_0 and $\alpha_i, i \in \{1,\ldots,6\}$ as

$$\alpha_0 = \sum_{i=1}^{6} \alpha_i. \tag{3.28}$$

$$\alpha_i = 2h \frac{\sigma_0 \sigma_i}{\sigma_0 + \sigma_i}, \tag{3.29}$$

where

σ_0 is conductivity within G_0
σ_i is conductivity in the ith region outside G_0

Based on this, a finite difference approximation of (3.5) will take the form

$$\sum_{i=1}^{6} \alpha_i V_{Q_i} - \alpha_0 V_P = I_P, \tag{3.30}$$

where $I_P = \iiint_{G_0} -I\delta(x-x_1)\delta(y-y_1)\delta(z-z_1) + I\delta(x-x_2)\delta(y-y_2)\delta(z-z_2)\,dxdydz$. The interpretation of this is as follows: Let $r=(x,y,z)$, $r_1=(x_1,y_1,z_1)$, and $r_2=(x_2,y_2,z_2)$, where r is situated on the line joining r_1 and r_2 and somewhere in between the two. If r_1 and r_2 are two poles of a dipole, then one has current I and the other $-I$. α_i corresponds to conductance between P and Q_i and has the dimension of conductance. I_P corresponds to current and takes the value I or $-I$ at Q_i and it has the dimension of current.

If P is at the center of mass of the box G_0, then all diagonals of G_0 must pass through P, and h is the distance from P to Q_i for $i \in \{1,\ldots,6\}$. If Q_5 is r_1 then Q_6 is r_2 and $\{-I\delta(x-x_1)\delta(y-y_1)\delta(z-z_1) + I\delta(x-x_2)\delta(y-y_2)\delta(z-z_2)\}dxdydz$ is to be calculated on the line joining Q_5 and Q_6. When calculations across all the three lines in Figure 3.6 are done and box G_0 is small enough, we can say that $I_P = \iiint_{G_0} -I\delta(x-x_1)\delta(y-y_1)\delta(z-z_1) + I\delta(x-x_2)\delta(y-y_2)\delta(z-z_2)\,dxdydz$ has been performed.

If there are n number of cubic grids, as shown in Figure 3.6, then there are n equations of the form (3.30) alongside n equations of the forms (3.28) and (3.29). n equations of the form (3.30) can be written in matrix form as

$$\mathbf{AV} = \mathbf{I}, \tag{3.31}$$

where

A is a $n \times n$ matrix with elements of the form α_i (known)
V is an n-dimensional vector of voltages of the form $V_{Q_i} - V_P$ (V_P is known)
I is a n-dimensional vector of currents of the form I_P (known)

At each stage, V_{Qi} will have to be calculated. At the time of computation, (3.31) is treated as a large sparse matrix (n may be of the order of several thousand). The solution is iterative with n number of iterations. The matrix and vectors in (3.31) are sparse. In each iteration, they have nonzero entries only corresponding to the value of one particular cubic element of the form of Figure 3.6. All other entries are zero.

3.1.3.2 aFDM

In anisotropic tissues, the conductivity tensor can vary between neighboring elements. There are two ways to have anisotropy. In general, anisotropy can be in any direction. The ideal mathematical tool to model this anisotropic conductivity is a tensor. Using tensor transformations, the matrix representation of the conductivity tensor can be deduced. In a more specific case, the directions of anisotropy are limited along the axes of the coordinate system of the head model. This is called *orthotropic* anisotropy. The conductivity tensor in such a case is given by Equation 3.33 for three dimensions. In the general case, the off-diagonal entries of the conductivity tensor matrix will be nonzero.

When electrical potential crosses a boundary separating two regions with different conductivity values, the potential cannot have a discontinuity on the boundary. This does not hold if one of the conductivities is zero (such as conductivity of air). This is called Dirichlet's boundary condition and is expressed as

$$V_1 = V_2, \tag{3.32}$$

where V_1 and V_2 are potentials on either side of the boundary. Now Equations 3.3, 3.4, and 3.32 together form a set of linear equations even for anisotropic media. Here, within each cubic grid (element), conductivity varies with respect to the direction of the current flow. If the coordinate system within the element coincides with the axes representing the principal directions of anisotropy, then the conductivity tensor at an element can be written as a diagonal matrix. The diagonal entries represent the conductivities in orthogonal directions. The matrix representation has to be transformed to the global Cartesian coordinate system of the head, the same for all elements. A rotation matrix is then required to transform the principal directions to a conductivity tensor in the Cartesian coordinate system. In the local coordinate system, the jth element can be written as follows:

$$\sigma^{j} = \begin{bmatrix} \sigma_1^j & 0 & 0 \\ 0 & \sigma_2^j & 0 \\ 0 & 0 & \sigma_3^j \end{bmatrix}, \tag{3.33}$$

where σ_i^j are the conductivities in the principal directions at element j ($i = 1$ along x axis, etc.). The matrix representation has to be transformed to a global Cartesian coordinate system of the head. Therefore a rotation matrix has to be applied. The matrix representation of the conductivity tensor at each element j in the Cartesian system of the head is then given by $\sigma_{head}^j = \mathbf{T}_j^T \sigma^j \mathbf{T}_j$, where $\mathbf{T}$ is a rotation transfer matrix from the local coordinate system to the global coordinate system.

Now let us go for the general case where anisotropy can be in any direction. A typical node 0 in the grid represents the interaction of eight neighboring cubic elements as shown in Figure 3.7. The conductivity within a cubic volume centering the node of interest (here it is node 0) has been discretized into eight subcubes with continuously inhomogeneous conductivity. The 18 neighboring points have been selected according to the scheme of Saleheen and Ng (1997). The rectilinear distance of the point 0 from each of the

18 neighboring points is less than or equal to 2 (Hallez et al., 2007), and therefore those 18 points have been chosen according to the nearest neighbor concept.

$$\sum_{i=1}^{18} a_i V_i = \left(\sum_{i=1}^{18} a_i \right) V_0, \tag{3.34}$$

where

V_0 is the potential at the point 0

V_i is the potential at the ith neighboring point (Saleheen and Ng, 1997)

a_i is a coefficient depending on the conductivity tensor of the corresponding element and the internode distance

a_i has been approximated for $i \in \{1, \ldots, 18\}$ in Saleheen and Ng (1997) as follows:

$$a_1 = \frac{1}{4\Delta x^2}\left[\sigma_{11(3)} + \sigma_{11(4)} + \sigma_{11(7)} + \sigma_{11(8)}\right],$$

$$a_2 = \frac{1}{4\Delta y^2}\left[\sigma_{22(1)} + \sigma_{22(4)} + \sigma_{22(5)} + \sigma_{22(8)}\right],$$

$$a_3 = \frac{1}{4\Delta x^2}\left[\sigma_{11(1)} + \sigma_{11(2)} + \sigma_{11(5)} + \sigma_{11(6)}\right],$$

$$a_4 = \frac{1}{4\Delta y^2}\left[\sigma_{22(2)} + \sigma_{22(3)} + \sigma_{22(6)} + \sigma_{22(7)}\right],$$

$$a_5 = \frac{1}{4\Delta x \Delta y}\left[\sigma_{12(4)} + \sigma_{12(8)}\right],$$

$$a_6 = -\frac{1}{4\Delta x \Delta y}\left[\sigma_{12(1)} + \sigma_{12(5)}\right],$$

$$a_7 = \frac{1}{4\Delta x \Delta y}\left[\sigma_{12(2)} + \sigma_{12(6)}\right],$$

$$a_8 = -\frac{1}{4\Delta x \Delta y}\left[\sigma_{12(3)} + \sigma_{12(7)}\right],$$

$$a_9 = \frac{1}{4\Delta z^2}\left[\sigma_{33(5)} + \sigma_{33(6)} + \sigma_{33(7)} + \sigma_{33(8)}\right],$$

$$a_{10} = \frac{1}{4\Delta z^2}\left[\sigma_{33(1)} + \sigma_{33(2)} + \sigma_{33(3)} + \sigma_{33(4)}\right],$$

$$a_{11} = \frac{1}{4\Delta y\Delta z}\left[\sigma_{23(5)} + \sigma_{23(8)}\right],$$

$$a_{12} = \frac{1}{4\Delta y\Delta z}\left[\sigma_{23(1)} + \sigma_{23(4)}\right],$$

$$a_{13} = \frac{1}{4\Delta y\Delta z}\left[\sigma_{23(2)} + \sigma_{23(3)}\right],$$

$$a_{14} = \frac{1}{4\Delta y\Delta z}\left[\sigma_{23(6)} + \sigma_{23(7)}\right],$$

$$a_{15} = \frac{1}{4\Delta x\Delta z}\left[\sigma_{13(7)} + \sigma_{13(8)}\right],$$

$$a_{16} = -\frac{1}{4\Delta x\Delta z}\left[\sigma_{13(5)} + \sigma_{13(6)}\right],$$

$$a_{17} = \frac{1}{4\Delta x\Delta z}\left[\sigma_{13(1)} + \sigma_{13(2)}\right],$$

$$a_{18} = -\frac{1}{4\Delta x\Delta z}\left[\sigma_{13(3)} + \sigma_{13(4)}\right],$$

where the subscript number in bracket is the node at which the conductivity or its derivative has been calculated (for an explanation, see Saleheen and Ng, 1997).

3.1.4 Comparison among Methods

We have discussed all four methods of solving the forward problem: BEM, FEM, iFDM, and aFDM. We present a summary of comparative studies among the four paradigms of forward problem solvers in Table 3.2.

TABLE 3.2

Comparison among Different EEG Forward Problem Solvers

	BEM	FEM	iFDM	aFDM
Position of computational points	Surface	Volume	Volume	Volume
Choice of computational points	Arbitrary	Arbitrary	Specific	Specific
Gain matrix	Full	Sparse	Sparse	Sparse
Number of compartments	Equal to the number of tissue layers	Equal to the number of computational elements	Equal to the number of computational elements	Equal to the number of computational elements
Tessellation	Not done	Done	Done	Done
Anisotropy handling	Not capable	Capable	Not capable	Capable
Solver	Direct	Iterative	Iterative	Iterative
Time complexity	Low	High	Higher than BEM but lower than FEM	Lower than FEM but higher than iFDM

3.2 Inverse Problem

In the previous section, we have seen how the head model is constructed, so that given a current source inside the brain we can determine the electric potential at the sensors on the scalp. Now we will do the opposite, that is, given the potential at the sensors on the scalp, we will determine their cortical sources. In other words we will 'invert' the gain matrix in some sense.

3.2.1 Weighted Minimum Norm Inverse

Here we describe a generalized form of the classical minimum norm inverse problem, from which we derive the minimum norm problem as a special case. The EEG inverse problem is inherently ill-posed, because the number of sources is usually higher than the number of sensors on the scalp. There are infinitely many different cortical source configurations possible for a single scalp electrode recording configuration. So, in general, no unique inverse source localization is possible for scalp recording. Therefore the problem is ill-posed. However, the situation can be improved by imposing constraints so that only one inverse source configuration is possible for a given scalp recording. The earliest such constraint is known as minimum norm inverse. Here we present a general formulation, namely, weighted minimum norm inverse. Minimum norm will follow when weights are unity. Figure 3.8 gives a geometric interpretation of how weighted minimum norm can lead to a

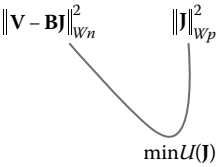


FIGURE 3.8
Geometric interpretation of weighted minimum norm inverse, where $U(\mathbf{J})=(1-\lambda)\|\mathbf{V}-\mathbf{BJ}\|^2_{Wn}+\lambda\|\mathbf{J}\|^2_{Wp}$, $0\le\lambda\le1$. When λ is very small, $U(\mathbf{J})=\|\mathbf{V}-\mathbf{BJ}\|^2_{Wn}+\lambda\|\mathbf{J}\|^2_{Wp}$. There is a unique value of $\mathbf{J}$ for which $U(\mathbf{J})$ will be minimum.

unique solution in terms of the source vector **J**. For simplicity, we consider distributed source model only, which can be generalized to dipole source model by considering three directional components, each will be treated like a distributed source model.

If the scalp potential vector is **V**, the gain matrix is **B**, and the additive noise vector is **E**, then the inverse problem for a single time point can be formulated as

$$\mathbf{V}=\mathbf{BJ}+\mathbf{E}. \tag{3.35}$$

In (3.35) we know **V** by measuring the scalp EEG. We also know **B** by solving the forward problem. The challenge of inverse problem is to find **J**. Let us define a quadratic energy function U as

$$U(\mathbf{J})=\|\mathbf{V}-\mathbf{BJ}\|^2_{Wn}+\lambda\|\mathbf{J}\|^2_{Wp}, \tag{3.36}$$

where $\|\mathbf{V}-\mathbf{BJ}\|^2_{Wn}+\lambda\|\mathbf{J}\|^2_{Wp}=\langle\mathbf{W}_n(\mathbf{V}-\mathbf{BJ}),\mathbf{W}_n(\mathbf{V}-\mathbf{BJ})\rangle+\lambda\langle\mathbf{W}_p\mathbf{J},\mathbf{W}_p\mathbf{J}\rangle$ and $\langle\ \rangle$ denotes inner product. $\mathbf{W}_n$ and $\mathbf{W}_p$ are weight matrices of order $n\times n$ and $p\times p$, respectively. λ is known as regularization parameter. Equation 3.36 is known as weighted minimum norm formulation. We will see in a moment how they are associated with **V**, **B**, and **J**. Figure 3.8 gives a geometric representation of U (a scalar-valued function of the vector variable **J**). Notice that $\|\mathbf{V}-\mathbf{BJ}\|^2_{Wn}=r$ is an ellipsoid whose variables are components of the source vector **J**, where r is a positive real number. Similarly, $\lambda\|\mathbf{J}\|^2_{Wp}=r'$, where r' is a positive real number, is another ellipsoid. So $U(\mathbf{J})=r+r'$ is an ellipsoid, which has both a unique minimum and a unique maximum value (point). The minimum point has been shown in Figure 3.8.

In (3.36) when $U(\mathbf{J})$ is optimum, $\nabla_{\mathbf{J}}U(\mathbf{J})=0$, where $\nabla_{\mathbf{J}}$ is the first derivative with respect to **J** (partial derivative with respect to each entry in **J**). This implies

$$\mathbf{J}=\mathbf{C}_p\mathbf{B}^T\left[\mathbf{BC}_p\mathbf{B}^T+\mathbf{C}_n\right]^{-1}\mathbf{V}, \tag{3.37}$$

where $\mathbf{C}_p=(\lambda\mathbf{W}_p^T\mathbf{W}_p)^{-1}$ and $\mathbf{C}_n=(\mathbf{W}_n^T\mathbf{W}_n)^{-1}$ (Mattout et al., 2006). The formula $\langle\mathbf{AB},\mathbf{C}\rangle=\langle\mathbf{B},\mathbf{A}^T\mathbf{C}\rangle$ ($\mathbf{A},\mathbf{B},\mathbf{C}$ are matrices of appropriate dimension) has been used to derive (3.37) from $\nabla_{\mathbf{J}}\left(\|\mathbf{V}-\mathbf{BJ}\|^2_{Wn}+\lambda\|\mathbf{J}\|^2_{Wp}\right)=0$. If we can

estimate $\mathbf{C}_n$ and $\mathbf{C}_p$, we will be able to estimate $\mathbf{J}$ from (3.37). Note that an optimum $U(\mathbf{J})$ will be achieved for a unique $\mathbf{J}$, which for the minimum $U(\mathbf{J})$ has been elaborated in Figure 3.8.

Both $\mathbf{C}_n$ and $\mathbf{C}_p$ are symmetric square matrices. A good candidate for $\mathbf{C}_n$ is the covariance matrix of the signals emanating out of n channels on the scalp. $\mathbf{C}_p$ is usually taken as an operator (p, the number of sources, is assumed to be known, so that the size of the operator is also known). For different operators, different inverse solutions will be produced. This is the way every EEG/MEG inverse solution fundamentally starts as a general form of minimum norm inverse solution, namely, the weighted minimum norm inverse solution. When $\mathbf{C}_p = \mathbf{I}_p$, we get the classical minimum norm inverse solution, where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

There are specific methods to choose the value of the *regularization parameter* λ in (3.36). Here we will mention only the L-curve method. According to this method, logarithm over two different components of (3.36), namely, $\|\mathbf{V} - \mathbf{BJ}\|^2_{Wn}$ and $\|\mathbf{J}\|^2_{Wp}$, is taken. $\log\|\mathbf{J}\|^2_{Wp}$ is plotted against $\log\|\mathbf{V} - \mathbf{BJ}\|^2_{Wn}$ and the plot gives an L-shaped curve. The abscissa value at the point of maximum curvature of this curve (the corner point on the L shape) is taken to be the optimum value of λ (Grech et al., 2008).

Source localization by minimum norm inverse can be further improved with additional constraints. For example, it is possible to improve the accuracy of source localization by phase synchronization and signal power (Majumdar, 2009). The basic idea is that in the channels adjacent to a dominating source, signal power will be high and also the EEG collected through those channels will have higher degree of phase synchronization, because all the channels are supposed to contain a significant amount of signal from the same source.

3.2.2 MUSIC

MUSIC stands for MUltiple SIgnal Classification (Schmidt, 1986), which is used to describe techniques for localization of sources of multiple waveforms arriving at an array of sensors. The fundamental assumption here is that, each source is generating a signal which is independent as time series from a signal generated by any of the other sources, that is, pair wise correlation coefficient between the signal from the first source and that from any other source is zero. Another assumption is signals are all noisy and the noises are independent from the signals. Independent, pure signals make the signal subspace, where they remain superimposed on each other. They are to be separated out. We can represent the digitized scalp EEG as an $n \times T$ matrix $\mathbf{V}$, where n is the number of channels and T is the number of time points in the signal from each channel. We can write

$$\mathbf{V} = \mathbf{AJ} + \mathbf{E}, \tag{3.38}$$

where $\mathbf{A} = \mathbf{A}[\mathbf{a}(\mathbf{r}_1, \mathbf{u}_1), \ldots, \mathbf{a}(\mathbf{r}_i, \mathbf{u}_i), \ldots, \mathbf{a}(\mathbf{r}_p, \mathbf{u}_p)]$, $\mathbf{r}_i$ is the location vector of the ith dipole, $\mathbf{u}_i$ is the orientation vector of the ith dipole, $\mathbf{a}(\mathbf{r}_i, \mathbf{u}_i)$ is $n \times 1$ column

vector $[a_1(\mathbf{r}_i, \mathbf{u}_i), \ldots, a_j(\mathbf{r}_i, \mathbf{u}_i), \ldots, a_n(\mathbf{r}_i, \mathbf{u}_i)]^T$, $a_j(\mathbf{r}_i, \mathbf{u}_i)$ is the effect of the ith dipole on the jth channel, p is the number of dipoles, and $\mathbf{E}$ is the additive noise matrix of dimension $n \times T$. The first step of MUSIC is to create a square matrix out of $\mathbf{V}$ so that it can be decomposed into one of the matrix forms—canonical or standard—where its eigenvector space can be separated into independent basis vector components, which is the age-old technique for the separation of linear mixture into its independent components (sometimes called 'sources'). This is exactly the case in PCA and ICA. Now begins the mathematical trick in MUSIC.

$$E\left(\mathbf{V}\mathbf{V}^T\right) = \left(\mathbf{A}\mathbf{J}\right)\left(\mathbf{A}\mathbf{J}\right)^T + E\left(\mathbf{E}\mathbf{E}^T\right), \tag{3.39}$$

for $\mathbf{AJ}$ is the deterministic part of $\mathbf{V}$. Let the rank of the square matrix $(\mathbf{AJ})(\mathbf{AJ})^T$ be r. Then by singular value decomposition (SVD), we can write

$$\left(\mathbf{A}\mathbf{J}\right)\left(\mathbf{A}\mathbf{J}\right)^T = \Phi_S \Lambda_S \Phi_S^T, \tag{3.40}$$

where

Φ_S is $n \times r$ eigenvector matrix
Λ_S is $r \times r$ diagonal matrix of eigenvalues of $(\mathbf{AJ})(\mathbf{AJ})^T$

Now here is the most crucial step of MUSIC, which needs to be intuitively understood. $\mathbf{A}$ can be thought of as a matrix of 'signals' containing only information about position and orientation of the dipoles. Φ_S is the matrix of eigenvectors of $\mathbf{V}$ minus the noise part, that is, Φ_S is the matrix of transformed 'signals' such that the separation between the 'signals' is maximum (eigen space decomposition of the signal subspace). If this is clear, then it is obvious that where there is a dipole source for signals in $\mathbf{V}$ the correlation between the corresponding row vector of $\mathbf{A}$, say $[a_1(\mathbf{r}_i, \mathbf{u}_i), \ldots, a_j(\mathbf{r}_i, \mathbf{u}_i), \ldots, a_n(\mathbf{a}_i, \mathbf{u}_i)]$, will have a high value c_i with a row vector of Φ_S, say at the jth row. In fact, we will have to take the highest correlation value between $[a_1(\mathbf{r}_i, \mathbf{u}_i), \ldots, a_j(\mathbf{r}_i, \mathbf{u}_i), \ldots, a_n(\mathbf{a}_i, \mathbf{u}_i)]$ and the jth row where $j \in \{1, \ldots, i-1, i+1, \ldots, n\}$. This is expressed as subspace correlation $(\mathbf{A}, \Phi_S)$ and denoted as *subcorr* $(\mathbf{A}, \Phi_S)$ (Mosher and Lehay, 1998; note that in this paper inner product has been used in place of statistical correlation, but here statistical correlation is also quite appropriate and convenient to compute). Repeat this for all row vectors of $\mathbf{A}$ and we will get a sequence $\{c_1, \ldots, c_r\}$ with $0 < c_i \leq 1$ for all $i \in \{1, \ldots, r\}$. Now, form the sequence $\left\{1/\sqrt{1-c_i^2}\right\}_{i=1}^{r}$. If these values are plotted, as $c_i \to 1$, $1/\sqrt{1-c_i^2} \to \infty$, we will get a very high spike and this will indicate that the ith source is a significant source dipole. For this we can presume in the beginning that dipole sources are placed very closely and uniformly across the cortical surface. We can start with a large number of dipoles spaced closely and uniformly across the cortical surface. We retain as sources of interest only those for which c_is

are above a threshold close to 1. This step needs to be repeated at each iteration with fewer number of dipole sources to begin with. If one particular dipole source, say the *i*th, shows different high values of $1/\sqrt{1-c_i^2}$, in subsequent iterations that source must be due to a rotating dipole.

The parameters $\mathbf{r}_i$ and $\mathbf{u}_i$ in $\mathbf{A}$ are estimated by applying the least squares method on the expression $\{\hat{\mathbf{r}}_i, \hat{\mathbf{u}}_i\}_{ls} = \arg\min \|\mathbf{V} - \mathbf{A}\mathbf{A}^{\updownarrow}\mathbf{J}\|^2$ over the argument $\mathbf{J}$, where $\mathbf{A}^{\updownarrow}$ is a Moore–Penrose pseudo inverse of $\mathbf{A}$, the subscript *ls* stands for least square.

3.2.3 R-MUSIC

Now we describe Recursive MUSIC, or R-MUSIC, algorithm (Mosher and Lehay, 1998), which is an improved version of the classical MUSIC algorithm for source localization, as described in Section 3.2.2.

Design a sufficiently dense grid in the volume of interest, and at each grid point $\mathbf{r}_i$ a single dipole gain matrix $\mathbf{G}(\mathbf{r}_i)$ is formed. This is accomplished by the forward model, which must be calculated before the inverse calculation.

For *index* from 1 to rank *r*

1. Let $\hat{\mathbf{A}} = \left[\mathbf{a}_1, \ldots, \mathbf{a}_{(index-1)}\right]$ be the model extracted as of the previous loop ($\hat{\mathbf{A}}$ is a null matrix for the first loop).
2. Form all one dipolar topographies (for each possible single dipole position), form all two dipolar topographies (for each possible two dipole positions), form all three dipolar topographies, so forth. Obviously, the number of such topographies will blow up exponentially. So, make only a few random dipolar topographies $\{\mathbf{G}(\lambda_i)\}$, where $\lambda_i = [\mathbf{a}_{k_1}, \mathbf{a}_{k_2}, \ldots, \mathbf{a}_{k_i}]$, where $\mathbf{a}_{k_i}$ is a dipole location. $\mathbf{G}(\lambda_i)$ is a *i* dipolar topography or *i* dipolar forward model or *i* dipolar gain matrix.
3. For the sequence of grid points $\{\lambda_i\}$, form the grid model $\mathbf{M}_i = \left[\hat{\mathbf{A}}, \mathbf{G}(\lambda_i)\right]$ by concatenating the already extracted model $\hat{\mathbf{A}}$ with $\mathbf{G}(\lambda_i)$.
4. Calculate the set of subspace correlations, $\{c_1, c_2, \ldots\} = subcorr(\mathbf{M}_i, \Phi_S)$ as in Section 3.2.2.
5. Find the maximum over all sets of grid points λ_i for c_{index}. For example, for *index* = 2 find the maximum second subspace correlation, that is, correlation value only for two dipolar λ_i.
6. If c_{index} is a good correlation ($\geq$0.95 according to Mosher and Lehay (1998)) go to step 7, else go to step 8.
7. Since the correlation is high enough, we have found the best set of locations λ_{index}. Note that λ_{index} is a collection of location vectors of 'index' number of dipoles. The orientation of each of these dipoles can also be found, for which the reader is referred to Mosher and Lehay (1998).

8. Since the correlation is low, we do not have the best set of locations yet. So we can either (a) reorganize the locations of the dipoles that we already have or (b) increase the number of dipoles. After (a) and/or (b) go back to 2.
9. Increase the *index* by 1 and repeat the steps 1 to 8 unless all dipolar topographies are exhausted or there is no significant improvement in correlation for successive topographies.

Note that the R-MUSIC described here is different from the one suggested by Mosher and Lehay (1998). Although this version will be easier to implement, computational complexity will remain the same as the original R-MUSIC. In most practical situations, this complexity will be prohibitive (the head may contain 7,000–25,000 dipoles depending on the model) and therefore R-MUSIC is not a useful algorithm unless we are interested in only a few dipoles, possibly specific to a particular task performance by the subject, whose EEG is being studied for task-specific localization of activations.

3.2.4 sLORETA

Standardized LOw Resolution (brain) Electromagnetic TomogrAphy (sLORETA) is claimed to be the most effective solution to the EEG/MEG inverse problem in a three-dimensional realistic head model (Pascual-Marqui, 2002). It starts by introducing a new parameter c in Equation 3.35 as

$$\mathbf{V} = \mathbf{BJ} + c\mathbf{I}_1, \tag{3.41}$$

where $\mathbf{I}_1$ is a vector of all ones, whose dimension is equal to that of $\mathbf{V}$. Note that additive noise $\mathbf{E}$ has been removed in (3.41). Therefore the signals must be noise free as far as possible for the accuracy of the localized cortical sources. c is an arbitrary constant that embodies the fact that the electric potential is determined up to an arbitrary constant. The parameter c allows the use of any reference for the lead field and the measurements (Pascual-Marqui, 2002). Now (3.36) becomes

$$U(\mathbf{J}) = \|\mathbf{V} - \mathbf{BJ} - c\mathbf{I}_1\|^2 + \lambda\|\mathbf{J}\|^2. \tag{3.42}$$

$U(\mathbf{J})$ is to be minimized with respect to c and $\mathbf{J}$ for given $\mathbf{B}$, $\mathbf{V}$, and λ. The expected solution of this problem is

$$\hat{\mathbf{J}} = \mathbf{TV}, \tag{3.43}$$

where $\hat{\mathbf{J}}$ is an estimation of $\mathbf{J}$ and

$$\mathbf{T} = \mathbf{B}^T\mathbf{H}\left[\mathbf{HBB}^T\mathbf{H} + \lambda\mathbf{H}\right]^+, \tag{3.44}$$

where

$$\mathbf{H} = \mathbf{I} - \mathbf{I}_1\mathbf{I}_1^T / \mathbf{I}_1^T\mathbf{I}_1, \tag{3.45}$$

I is the identity matrix of appropriate dimension. **H** is called the centering matrix. $\mathbf{A}^+$ denotes the Moore–Penrose inverse of the matrix **A**. It is assumed that the variance in $\hat{\mathbf{J}}$ is due to the variance in actual source vector **J**. Clearly **H** will be of the form

$$\begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{pmatrix}$$

where n is the number of scalp EEG electrodes.

References

Acar, Z. A. and S. Makeig, Effects of forward model errors on EEG source localization, *Brain Topogr.*, **26**: 378–396, 2013.

Geddes, L. A. and L. E. Baker, The specific resistance of biological material—A compendium of data for the biomedical engineer and physiologist, *Med. Biol. Eng.*, **5(3)**: 271–293, 1967.

Gencer, N. G. and C. E. Acar, Sensitivity of EEG and MEG measurements to tissue conductivity, *Phys. Med. Biol.*, **49**: 701–717, 2004.

Goncalves, S., J. de Munck, J. Verbunt, F. Bijma, R. Heethaar, and F. Lopes de Silva, In vivo measurements of the brain and skull resistivity using an EIT-based method and realistic model for the head, *IEEE Trans. Biomed. Eng.*, **50(6)**: 754–757, 2003.

Grech, R., T. Cassar, J. Muscat, K. P. Camilleri, S. G. Fabri, M. Zervakis, P. Xanthopolus, V. Sakkalis, and B. Venrumste, Review on solving of inverse problem in EEG source analysis, *J. NeuroEng. Rehab.*, **5**: 25, 2008.

Gutirrez, D., A. Nehorai, and C. H. Muravchik, Estimating brain conductivities and dipole source signals with EEG arrays, *IEEE Trans. Biomed. Eng.*, **51(12)**: 2113–2122, 2004.

Hallez, H., B. Venrumste, R. Grech, J. Muscat, W. De Clercq, A. Vergult, Y. D'Asseler et al., Review on solving the forward problem in EEG source analysis, *J. NeuroEng. Rehab.*, **4**: 46, 2007.

Kybic, J., M. Clerc, T. Abboud, O. Faugeras, R. Keriven, and T. Papadopoulo, A common formalism for integral formulations of the forward EEG problem, *IEEE Trans. Med. Imaging*, **24(1)**: 12–28, 2005.

Kybic, J., M. Clerc, O. Faugeras, R. Keriven, and T. Papadopoulo, Generalized head model for MEG/EEG: Boundary element method beyond nested volumes, *Phys. Med. Biol.*, **51**: 1333–1346, 2006.

Lai, Y., W. van Drongelen, L. Ding, K. E. Hecox, V. L. Towle, D. M. Frim, and B. He, Estimation of in vivo human brain-to-skull conductivity from simultaneous extra- and intra-cranial electrical potential recording, *Clin. Neurophysiol.*, **116(2)**: 456–465, 2005.

Majumdar, K., Constraining minimum norm inverse by phase synchronization and signal power of the scalp EEG channels, *IEEE Trans. Biomed. Eng.*, **56(4)**: 1228–1235, 2009.

Mattout, J., C. Philips, W. D. Penny, M. D. Rugg, and K. J. Friston, MEG source localization under multiple constraints: An extended Bayesian framework, *NeuroImage*, **30**: 753–767, 2006.

Mosher, J. C. and R. M. Lehay, Recursive MUSIC: A framework for EEG and MEG source localization, *IEEE Trans. Biomed. Eng.*, **45(11)**: 1342–1354, 1998.

Oostendorp, T., J. Delbeke, and D. Stegeman, The conductivity of the human skull: Results of in vivo and in vitro measurements, *IEEE Trans. Biomed. Eng.*, **47(11)**: 1487–1492, 2000.

OpenMEEG 2.1, A free software for solving EEG forward problem, 2011, available at https://gforge.inria.fr/frs/?group_id=435 (accessed April 3, 2014).

Pascual-Marqui, R. D., Standardized low resolution brain electromagnetic tomography (sLORETA): Technical detail, *Methods Find. Exp. Clin. Pharmacol.*, **24(Suppl D)**: 5–12, 2002.

Saleheen, H. I. and Ng, K. T., New finite difference formulations for general inhomogeneous anisotropic bioelectric problems, *IEEE Trans. Biomed. Eng.*, **44(9)**: 800–809, 1997.

Schmidt, R. O., Multiple emitter location and signal parameter estimation, *IEEE Trans. Antenna Prop.*, **34(3)**: 276–280, 1986.

Yan, Y., P. L. Nunez, and R. T. Hart, Finite-element model of the human head: Scalp potential due to dipole sources, *Med. Biol. Eng. Comput.*, **29**: 475–481, 1991.

4

Event-Related Potential

We have seen in Chapter 1 how electrical activities of ensembles of neurons and synapses give rise to measurable electrical potential on the scalp (from −100 to +100 microvolts (μV), which is usually amplified with an amplifier for better visualization and processing). When the scalp potential is evoked in response to an event or a stimulus present in the environment, the evoked potential is called event-related potential (ERP). Of course, in a scalp EEG, the ERP may be embedded in artifacts and neuronal activities not evoked by the event of interest.

4.1 Plotting ERP Data

Usually, for a good ERP, the EEG data is averaged across the trials. The assumption is that noise is Gaussian in nature and therefore almost equally has positive and negative values, which, upon adding up cancel, leaving the ERP more prominent. The ERP is assumed to be similar for the same stimulus, and therefore only adds up to become more prominent (Figure 4.1). However, it cannot be utilized in single-trial EEG analysis. ERP contains both, the potential directly evoked in the brain by sensing the stimulus (evoked potential) and the electric potential generated in the brain by the cognitive processes associated with the stimulus. The visual analysis of ERP is predominantly popular among cognitive psychologists, and for this an ERP signal is divided into positive (P) and negative (N) components (Figure 4.1).

Although all sensory perceptions namely, visual, auditory, olfactory, somatosensory, and gustatory generate the ERP, the most widely studied among them is visual ERP. Some prominent visual ERPs are P300 (on an average the large positive potential observed in the occipital lobe 300 ms after the onset of a visual stimulus), N100 (on an average the negative potential observed in the occipital lobe 100 ms after the onset of a visual stimulus), etc. (also known as P3, N1, etc., respectively). However, in this treatise our focus will be on quantitative analysis of the ERP rather than visual analysis. This almost entirely relies on the enhanced magnitude of the ERP compared to the background EEG (this also offers the biggest clue to ERP identification

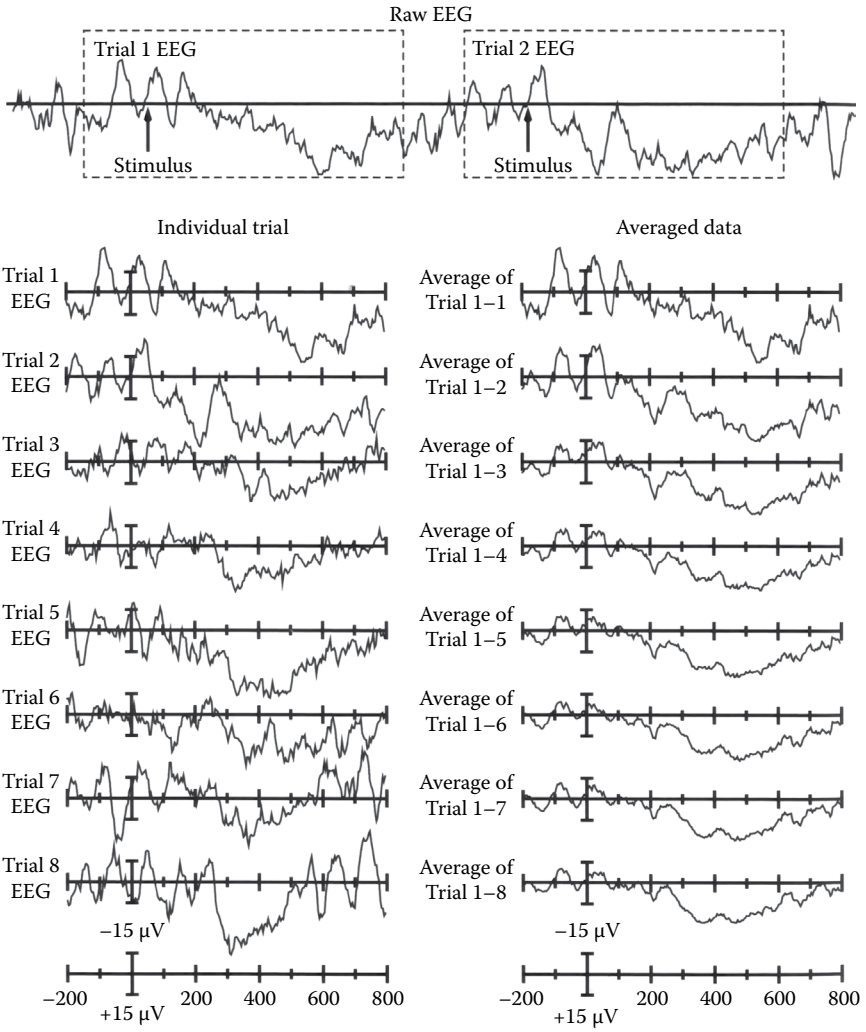


FIGURE 4.1
Event-related potential single trial (left) and averaged across trials (right). Parts falling above the thick horizontal line belong to the class of P component, and those falling below that line fall among the class of N components. (Adapted from motorbehaviour.wordpress.com.)

in visual analysis). PCA and ICA discussed in Chapter 2 are very useful tools for automatic ERP identification in EEG recordings. Magnitude difference can be plotted in a visually appealing way by means of topographical plots or simply topoplots (Figure 4.2). EEGLAB offers a host of options for two- and three-dimensional topographical plots of EEG data for identifying the ERP segments. The plot is available in the EEGLAB window itself. In order

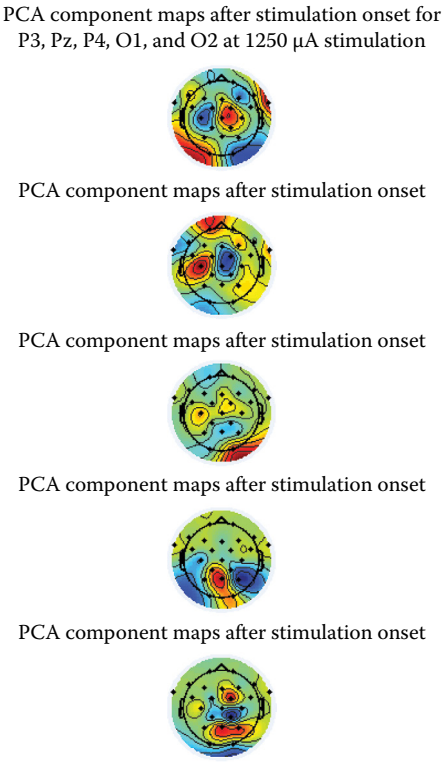


FIGURE 4.2
Topoplot during transcranial alternating current stimulation at the maximum intensity of 1250 μA. The first five principal components have been shown in the decreasing order from top to bottom. They are associated with the channels P3, Pz, P4, O1, and O2, respectively. Red is maximum positive and blue is minimum negative. (Courtesy of Christoph Herrmann, University of Oldenburg, Oldenburg, Germany.)

to generate the topoplots the channel location needs to be fixed. This in the EEGLAB comes in the form of a .elp file.

At this point let us familiarize ourselves with another concept very closely associated with the brain dynamics, which goes beyond simple ERPs. This is called event-related spectral perturbation (ERSP). Cortical regions are specialized for specific tasks. These regions oscillate either coherently or incoherently among themselves. This oscillation is spontaneous and does not depend on specific stimuli. But once a stimulus is presented, a distinct change in the coherent and incoherent oscillation pattern is observed (if coherent, it is called event-related synchronization [ERS], and if incoherent, it is called event-related desynchronization [ERD]). The change of oscillation patterns with respect to the baseline oscillation in the presence of a stimulus is termed as ERSP (Makeig, 1993).

It is calculated by time-frequency analysis with the help of windowed fast Fourier transform (FFT) in the following steps:

1. Choose a sliding temporal window with a suitable size and overlap for the selected EEG epoch.
2. Perform FFT in the window and determine the power spectral density for a frequency bin size of F_n (1 Hz or slightly larger).
3. Smoothen by a moving average window of suitable size in order to eliminate the random jitters.
4. All smooth window-wise spectral estimates obtained by 2 and 3 are normalized by dividing by the spectral estimate of the first window (EEG epoch to be selected in such a way prior to 1 that this first estimate will correspond to a prestimulus or baseline spectrum). It can be expressed in the decibel (dB) form by taking $20\log_{10}$ of that ratio.

In Figure 4.3, the ERSP of a subject in a single EEG channel on the occipital lobe has been shown during visual perception, followed by a key-pressing task. Notice that the high spectral perturbation in the upper alpha (10–12 Hz) and lower beta (13–20 Hz) ranges more than 1000 ms after the stimulus onset, that is, roughly at the time of key pressing. This is predominantly due to the action planning and execution, which cannot be captured by mere ERP. ERSP is shown using the time–frequency color plot. The upper left marginal panel presents the mean spectrum during the baseline period. The marginal panel under the ERSP image shows the maximum (green) and minimum (blue) ERSP values relative to the baseline power at each frequency.

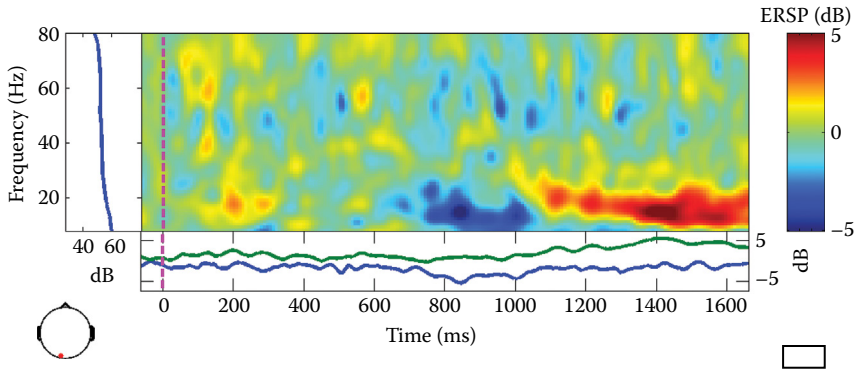


FIGURE 4.3
ERSP after a visual stimulus onset (vertical dashed line at 0 time point). The task involves key pressing followed by the stimulus presentation. The channel position in the occipital lobe is shown in the left-bottom corner. The corresponding ERP is shown by a green line. The ERSP is shown by a blue line in both frequency and time domains (by averaging across the frequencies).

4.2 Measuring ERP Amplitudes

Most ERP studies concentrate on the amplitudes of one or more ERP components although amplitudes are dependent on the reference channel (the average reference has been advocated by Murray et al. (2008)). There are two common ways to measure ERP amplitudes. The most common method is to fix a time window, and, for each waveform being measured, find the maximum amplitude in that time window. This is called *peak amplitude measure*. Instead of the maximum amplitude when the average amplitude of a waveform in the window is measured it is called *mean amplitude measure*, which is the second-most common way to measure the ERP amplitude (Luck, 2005). It is assumed that the ERP amplitude before presentation of the stimulus is the baseline amplitude and is subtracted from the measured ERP amplitude. This subtracted value is considered to be the actual measure of the ERP amplitude. For this purpose the mean amplitude for a period immediately preceding the stimulus onset is taken as the baseline amplitude to be subtracted.

4.2.1 Peak Amplitude

Peaks in a signal can be detected in numerous different ways. The easiest among them is taking a triplet of successive points and seeing if the midpoint has a higher amplitude than any of the points on either side. Clearly, by this method, the detected peaks will be highly noisy. Here is a simple method for detecting peaks and negative peaks or troughs with 100% accuracy (however, it is prone to false positives and therefore amplitude thresholding along with inclusion of other features may be necessary). Low-frequency and low-amplitude noise are automatically suppressed (Majumdar et al., 2014).

The peaks and troughs in a digital signal s will appear as in Figure 4.4. Let the coordinates of B be $(m, s(m))$, those of A be $(m-1, s(m-1))$, and of C be

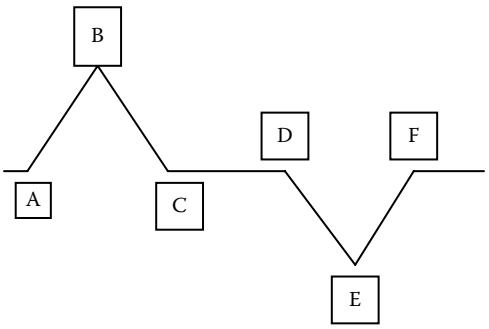


FIGURE 4.4
A peak (B) and a trough (E) in a digital signal.

$(m+1, s(m+1))$. Clearly, $s(m)-s(m-1)>0$, $s(m+1)-s(m)<0$. Note that $s''(m)=s(m+1)-s(m)-(s(m)-s(m-1))<0$. So $P(m-)=s''(m)(s(m)-s(m-1))<0$, and $P(m+)=s''(m)(s(m+1)-s(m))>0$. Therefore in the case of digital signals a peak at m can be identified by the property

$$P(m-)<0 \quad \text{and} \quad P(m+)>0 \quad \text{and} \quad s''(m)<0. \tag{4.1}$$

Similarly, the trough at E in Figure 4.1 with coordinates $(k, s(k))$ can be identified by the property

$$P(k-)<0 \quad \text{and} \quad P(k+)>0 \quad \text{and} \quad s''(k)>0. \tag{4.2}$$

Note that a difference operation works as a high-pass filter. The operator P will therefore work as a high-pass filter, allowing the high frequency noise to pass through. To avoid the former, a low-pass filter should be used before applying P. Regarding the latter, note that we are not using the signal morphology after being operated by P. We are only using the sign of P and s'' to identify a peak (Equation 4.1) or a trough (Equation 4.2), which remains true for peaks and troughs irrespective of their shape, amplitude, or any other morphological properties. Therefore, P acting as a high-pass filter does not affect the detection performance. It is clear that (4.1) will detect all peaks. For specific peaks, a suitable threshold may be used. A standard rule for setting a threshold is taking a window and then setting the threshold as mean + k times standard deviation of the signal within the window, where k is a nonnegative real number (usually from 0 to 4).

For peak amplitude measure, electrode location (e.g., visual ERP P300 is likely to occur most prominently in the electrodes placed over the occipital lobe) and peak latency (see the next section) after the stimulus onset are the two most important aspects to take into consideration. For example, P300 may occur 250 ms after the onset of stimulus in some trial and 350 ms after in some other for the same subject. Since the peak amplitude is the amplitude at a single time point, where the peak occurs, it is more vulnerable to noise and measurement error compared to mean amplitude measure.

The EEG signals collected from the subjects, during stimulus presentation, often contain hundreds of trials. Even after preprocessing, the signal strength across all the channels varies significantly from one trial to another leading to large differences in the peak amplitude across the trials. This variation is quite random across the trials. If the scalp EEG signal strength across all the channels in a trial is sufficiently weak, the ERP amplitude is likely to be faint and unreliable. To ensure adequate signal

strength in the trial under experimentation the following straightforward algorithm can be used (Majumdar, 2009):

Proc(trial_select)

1. **Identify the time interval;**
2. **Identify the channels (default all channels);**
3. **Select trials serially {**
   a) **During the trial sum the square of the values of the discrete signal across the time interval at each selected channel;**
   b) **Take the mean of the values obtained in 3.a) across all the selected channels.**

   }
4. **Sort the values obtained in 3 in decreasing order across all the trials.**
5. **Select as many trials as required from the top (high valued);**

4.2.2 Mean Amplitude

The difference operation works as a high-pass filter, windowed averaging being a kind of smoothing that works as a low-pass filter to eliminate high-frequency noise. Since high-frequency components (such as the gamma band EEG in the frontal lobe during planning or decision-making) are usually not associated with the ERP, the moving window mean amplitude measure can safely be used for ERP amplitude measure for wide-ranging tasks.

The window size for mean amplitude measure should be smaller compared to peak amplitude measure. This will give a sharper resolution of the signal and a more realistic measure of ERP amplitude. However, the window should not be narrower than 40 ms (Luck, 2005, p. 234).

4.3 Measuring ERP Latencies

ERP latency after the stimulus onset is strongly dependent on ERP waveform identification, which in turn is dependent on amplitude measure.

4.3.1 Peak Latency

Peak latency measure is tightly coupled with peak amplitude measure and therefore has all the drawbacks of the latter. However, there are not too many good alternative methods either. So peak latency measure can be employed

with the following precautions: (1) filter out the high-frequency noise in the EEG; (2) rather than taking the maximum peak alone, take other local peaks also (possibly with some threshold), because the maximum peak may not always be due to an ERP waveform; (3) when different waveforms are compared they must have a similar noise level.

4.3.2 Fractional Area Latency

According to this method, a temporal window has to be fixed (which is quite subjective). In Figure 4.5, the window has been taken from 17 to 48 units of time after the stimulus onset. Usually the window is taken to encompass the EEG where the ERP visually looks most prominent (one more subjectivity). Then the area under the curve within that window is measured (Figure 4.5). Now a time point is selected such that a predetermined fraction of the area, say p, falls on one side of that point and the fraction $1 - p$ of the area falls on the other side. The value of this time point is taken to be the *fractional area latency* of the ERP. Most often, p is taken as 0.5, because it becomes less susceptible to noise. In Figure 4.5, the fractional area latency is 32.5 units of time.

We have seen that both peak latency measure and fractional area latency measure are error prone. Fortunately, in many instances we do not need to measure the absolute latency, but rather difference of latency between two different stimulus conditions. For this, 50% fractional area latency measure

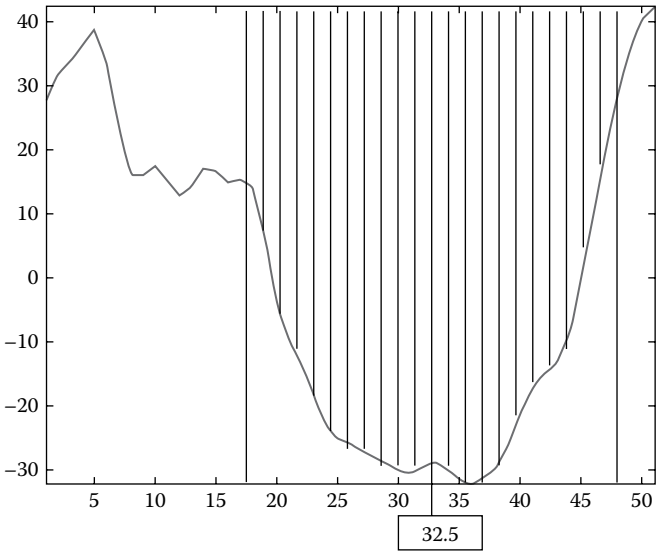


FIGURE 4.5
Time vs. amplitude plot of ERP (the curve). The (time) window is from 17 to 48. The shaded region is the area under the curve. The area under the curve is divided into two equal parts at the time point 32.5.

($p = 0.5$) works satisfactorily in many cases (Luck, 2005, p. 241) despite the subjective errors mentioned above.

4.4 Analyses of ERP

Once the ERP amplitude and latency are measured, they need to be analyzed for an interpretation of the underlying brain activities. A plethora of computational approaches are employed for this purpose (Murray et al., 2008). The main objective, however, is to see if ERP amplitude and latency in different parts of the scalp vary across the presented stimulus conditions. The oldest and the most popular analysis in this respect is analysis of variance (ANOVA).

4.4.1 ANOVA in ERP

Let us start with a concrete example. Consider a simple experiment, where the effect of sleep deprivation on reaction time is to be examined (Gonzalez, 2009). There are 32 subjects, who are to be allocated into 4 groups—A_1 (12-hour sleep deprivation), A_2 (24-hour sleep deprivation), A_3 (36-hour sleep deprivation), and A_4 (48-hour sleep deprivation). There are 8 subjects in each group. Table 4.1 contains the details of the response time (in seconds) for each subject. The hypothesis (known as alternate hypothesis) that sleep deprivation has had an effect on the response time is to be tested against the null hypothesis that there was no effect of sleep deprivation on the response time.

TABLE 4.1

Response Time for Subjects with Sleep Deprivation

	A_1 12 Hour	A_2 24 Hour	A_3 36 Hour	A_4 48 Hour
	20	21	25	26
	20	20	23	27
	17	21	22	24
	19	22	23	27
	20	20	21	25
	19	20	22	28
	21	23	22	26
	19	19	23	27
Group mean	19.38	20.75	22.63	26.25
Group std.	1.19	1.28	1.189	1.28

Note: Response time in seconds for the subjects appears in the boxes.

Let each data point in Table 4.1 be denoted by Y_{ij}, in which i denotes the group the subject belongs to, and j denotes that the subject is the jth person of the group. Obviously, if we do not know how much sleep deprivation a particular subject had, our best prediction of his or her reaction time would be the grand mean. It is not very precise, but if it is the only thing we have, we can live with it. This simple model of an individual data point can be denoted as follows:

$$Y_{ij} = \mu + \varepsilon_{ij}, \tag{4.3}$$

where

μ is the grand mean of all 32 response times of 32 subjects
ε_{ij} is the error or deviation from the grand mean

The error term is normally distributed with mean 0 and variance σ^2. Now, if we know the amount of sleep deprivation for individual subject then we can improve our hypothesis by using the group mean instead of the grand mean. The group mean can itself be considered to be a deviation from the grand mean. The ith group mean $\bar{Y}_i$ can be written as

$$\bar{Y}_i = \mu + \alpha_i, \tag{4.4}$$

where α_i is the deviation of the ith group mean from the grand mean. Now if we write Y_{ij} as

$$Y_{ij} = \bar{Y}_i + \varepsilon_{ij}, \tag{4.5}$$

ε_{ij} is the error or adjustment that has to be made with the ith group mean $\bar{Y}_i$ in order to get the individual data of jth member of the ith group Y_{ij}. Combining (4.4) and (4.5) we get a more refined form of (4.3) as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}. \tag{4.6}$$

Equation 4.6 is called the *full model* of each data value (e.g., each individual entry in Table 4.1). Notice that in (4.3) ε_{ij} is the adjustment with the grand mean μ to get Y_{ij}, but in (4.6) ε_{ij} is the adjustment with the group mean $\bar{Y}_i$ in order to get Y_{ij}. Equation 4.6 is one fundamental constraint of ANOVA namely, if errors or adjustments are not additive in a data set ANOVA cannot be used on that data set. Also, the distribution of all the data sets must be normal (Gaussian) with equal variance, which is a very strict condition indeed.

Going back to Table 4.1, if the mean of all the groups $A_1, \ldots, A_4$ are equal we can say that sleep deprivation does not have any effect on response time according to the current data. ANOVA takes as the null hypothesis the assertion that all the groups have equal mean. Note that if all the group means are equal, then they all are equal to the grand mean. In view of (4.6) an alternative way of stating this is $\alpha_i = 0$, for all i. The next question is how to relate this to 'analysis of variance'?

The variance of the entire data set is given by

$$\frac{\sum_i \sum_j \left(Y_{ij} - \mu\right)^2}{N-1} = \frac{SST}{DF},$$

where N is the total number of entries in the data set, SST stands for sum of square total, and DF stands for degrees of freedom.

$$SST = \sum_i \sum_j \left(Y_{ij} - \mu\right)^2 = \sum_i n_i \left(\bar{Y}_i - \mu\right)^2 + \sum_i \sum_j \left(Y_{ij} - \bar{Y}_i\right)^2 = SSB + SSW, \quad (4.7)$$

where

n_i is the number of entries in the ith group
SSB stands for sums of squares between groups
SSW stands for sum of squares within groups

Notice that if $\bar{Y}_j = \mu$ for all j, then from (4.7) we get $SSB = 0$ and $SST = SSW$. Only ANOVA can determine whether this is true.

Let us define two terms mean square between (MSB) and mean square within (MSW) as follows:

$$MSB = \frac{SSB}{DF_{SSB}} \quad \text{and} \quad MSW = \frac{SSW}{DF_{SSW}}, \quad (4.8)$$

where

DF_{SSB} is the number of degrees of freedom among the groups = number of groups – 1
DF_{SSW} is the number of degrees of freedom within a group = total number of data entries – number of entries in the group

For Table 4.1, $DF_{SSB} = 4 - 1 = 3$ and $DF_{SSW} = 32 - 4 = 28$. In ANOVA we are to measure the ratio $r = MSB/MSW$. If $r = 0$, then $SSB = 0$ and the null hypothesis is true. When $r > 0$, we should take recourse to some probability theory as described below.

If our original data entries Y_{ij} are normally distributed, then $Y_{ij} - \bar{Y}_i$ and $Y_{ij} - \mu$ are normally distributed. This means that $\left(Y_{ij} - \bar{Y}_i\right)^2$ and $(Y_{ij} - \mu)^2$ follow chi-square (denoted as χ^2) distribution. Also, the sum of chi-square-distributed variables follow chi-square distribution. Suppose two random variables X and Y follow chi-square distribution. Let X be sum of n variables, each following chi-square distribution and Y be sum of m variables, each following chi-square distribution. So X has $n-1$ degrees of freedom and Y has $m-1$ degrees of freedom. The ratio $\frac{X/(n-1)}{Y/(m-1)}$ will then follow F-distribution. Clearly, the quantity r follows F-distribution. Going back to Table 4.1, we can look for the value $F_{3,28;0.05}$ (the numerator has 3 degrees of freedom, the denominator has 28 degrees of freedom, and we are looking for the value where F-distribution is 0.05 so that we can reject the null hypothesis with 1 – 0.05 = 0.95 probability) in the F-distribution table. $F_{3,28}(r) = 0.05$ gives $r = 2.95$. But $r = MSB/MSW$

calculated from Table 4.1 is 46.56, which is greater than 2.95. Therefore, we can conclude that the probability of group means are not all equal is greater than 0.95. If some of the group means are equal and some others are not, this method will not be able to distinguish that from the situation where all group means are unequal. However, we already know that for ANOVA the data sets must be distributed normally. This condition can be relaxed when the sample size is sufficiently large, because by the central limit theorem the probability distribution of a sufficiently large sample is going to be a normal distribution.

For a typical application of ANOVA on ERP data, let us consider the following example from Luck (2005). In a visual odd-ball task a subject is shown X and O on a computer screen in a dark room. The subject has to press one button if it is X or another if it is O. Each symbol is shown for 100 ms with a gap of 1400 ms before the presentation of the next scene. Symbols are appearing randomly. In one set of trials, X is appearing 75% of the times and O 25% of the times. In the other set of trials, O is appearing 75% of the times and X 25% of the times. EEG collected during the trials from electrodes F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4 (see Figure 1.12).

Say we want to study P300 at Pz. Our objective is to see how the amplitude of P300 varies between frequent (75% occurrence) and rare (25% occurrence) stimuli. P300 amplitude is to be calculated as the mean amplitude between 300 and 800 ms after stimulus onset (Luck, 2005, p. 252). This is done through ANOVA, which says with how much certainty the amplitudes are different in the general case (i.e., in the actual population beyond the sample population that is being studied). Let us consider the trial block in which X appears 75% of the times and O 25% of the times. EEGs collected at Pz during all X trials are to be averaged across the trials from 300 to 800 ms after stimulus onset. The same procedure is followed for Pz during all O trials. Now we have two 500 ms long digitized EEG segments, each of which is a vector consisting of signal amplitude at each time point. These two vectors are two groups to be studied by ANOVA to examine if their mean values are same or not and with what amount of certainty for the actual population (actually the mean value for rare occurrence trials will be greater than the mean value of the frequent occurrence trials with probability 0.95 or more, which is the case for all odd-ball visual tasks).

Two crucial assumptions for the ANOVA are (1) each data set is distributed normally and (2) variances of all the data sets are equal to each other. Equality of variances can be tested through Bartlett's test. Skewness and kurtosis measures are the most elementary ways to test for normality of the data. There are many other tests for normality, for example, the Lilliefors test.

When normality assumption does not hold, nonparametric statistical tests are recommended. The Kruskal–Wallis test is one such widely applied test. Please see Gibbons (1971, pp. 198–201) for a very readable exposition.

4.4.2 MANOVA in ERP

Now if we want to study P300 for the same odd-ball task simultaneously in all the channels mentioned above, what do we do? One tempting way is to

perform ANOVA for every channel. But, then the power of the test will go down. What is the power of the test? Well, in every binary classification task (such as group means are all equal vs. not all group means are equal) there exist possibilities of two types of errors namely, type I error and type II error. Type I error is equivalently called false positive. It corresponds to rejection of the null hypothesis when the null hypothesis is true. Type II error is also known as false negative. It corresponds to rejection of the alternative hypothesis when the alternative hypothesis is true. In the above example, if we perform ANOVA for every channel separately, in the end result both type I and type II errors will be higher than when we perform 'ANOVA' on all the channels together. This is called multivariate ANOVA or MANOVA.

MANOVA starts with vectorization of the full model (4.6).

$$\mathbf{X}_{ij} = \mathbf{m} + \mathbf{a}_i + \mathbf{e}_{ij}, \tag{4.9}$$

where

- $\mathbf{X}_{ij}$ is the jth observation vector in the ith group (two important assumptions are (1) $\mathbf{X}_{ij}$ is distributed according to a multivariate normal distribution, and (2) covariance matrices of the groups are same)
- $\mathbf{m}$ is the global mean observation vector
- $\mathbf{a}_i$ is the 'adjustment' to made to $\mathbf{m}$ in order to get the mean for the ith group
- $\mathbf{e}_{ij}$ is the 'adjustment' to be made with the ith group mean in order to get the jth individual observation vector

Equation 4.9 can be rewritten as

$$\mathbf{X}_{ij} = \mathbf{m} + \left(\bar{\mathbf{X}}_i - \mathbf{m}\right) + \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right).$$

$$\left(\mathbf{X}_{ij} - \mathbf{m}\right)^2 = \left\{\left(\bar{\mathbf{X}}_i - \mathbf{m}\right) + \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}\right)\right\}^2. \tag{4.10}$$

Note that the product $(\mathbf{X}_{ij} - \mathbf{m})(\mathbf{X}_{ij} - \mathbf{m})'$, where $'$ denotes transpose, can be written as

$$\left(\mathbf{X}_{ij} - \mathbf{m}\right)\left(\mathbf{X}_{ij} - \mathbf{m}\right)' = \left[\left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right) + \left(\bar{\mathbf{X}}_i - \mathbf{m}\right)\right]\left[\left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right) + \left(\bar{\mathbf{X}}_i - \mathbf{m}\right)\right]'$$

$$= \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right)' + \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right)\left(\bar{\mathbf{X}}_i - \mathbf{m}\right)' + \left(\bar{\mathbf{X}}_i - \mathbf{m}\right)\left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right)' + \left(\bar{\mathbf{X}}_i - \mathbf{m}\right)\left(\bar{\mathbf{X}}_i - \mathbf{m}\right)'$$

Clearly, $\sum_j \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right)\left(\bar{\mathbf{X}}_i - \mathbf{m}\right)' = \left(\bar{\mathbf{X}}_i - \mathbf{m}\right)' \sum_j \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right) = 0.$ Similarly $\sum_j \left(\bar{\mathbf{X}}_i - \mathbf{m}\right)\left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right)' = 0$. With the help of this from (4.10) we get

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\mathbf{m}\right)\left(\mathbf{X}_{ij}-\mathbf{m}\right)' = \sum_{i=1}^{g} n_i\left(\bar{\mathbf{X}}_i-\mathbf{m}\right)\left(\bar{\mathbf{X}}_i-\mathbf{m}\right)' + \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\bar{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij}-\bar{\mathbf{X}}_i\right)', \tag{4.11}$$

where

g is the number of groups

n_i is the number of members in the ith group

The right-hand side of (4.11) can be written as $\mathbf{B}+\mathbf{W}$, where $\mathbf{B}=\sum_{i=1}^{g} n_i\left(\bar{\mathbf{X}}_i-\mathbf{m}\right)\left(\bar{\mathbf{X}}_i-\mathbf{m}\right)'$ and $\mathbf{W}=\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\bar{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij}-\bar{\mathbf{X}}_i\right)'$. Comparing (4.11) with (4.7) we observe that **B** and *SSB* are similar expressions in the sense they both represent the sum of the squares between groups. Similarly, **W** and *SSW* represent similar expressions in the sense that they both represent the sum of squares within groups. Equation 4.11 being a vector expression, we cannot in general obtain the *F*-distribution or some other elegant probability distribution for an expression involving **B** and **W** the way we were able do for ANOVA. Still we have several ways to ascertain whether $\bar{\mathbf{X}}_i=\mathbf{m}$ holds or not for all i. Here we will mention about only one of them namely, Wilk's lambda test. This is easier to perform and also more widely used than the others. For a discussion about other tests, see Johnson and Wichern (2002).

Wilk's lambda is denoted by Λ* and expressed by the formula $\Lambda^*=\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$. By (4.11) we can write

$$\Lambda^*=\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}=\frac{\left|\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\bar{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij}-\bar{\mathbf{X}}_i\right)'\right|}{\left|\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\mathbf{m}\right)\left(\mathbf{X}_{ij}-\mathbf{m}\right)'\right|}. \tag{4.12}$$

Here $|\mathbf{W}|$ is determinant of the matrix **W**. The ratio (4.12) is also called generalized variance (6-38 of Johnson and Wichern, 2002). There is another more popular expression of Wilk's lambda in terms of the eigenvalue of a product of **B** and **W**). We reject the null hypothesis $\bar{\mathbf{X}}_i=\mathbf{m}$ for all i, if the generalized variance or Wilk's lambda Λ* is too small. Its value always lies between 0 (means are different) and 1 (means are all equal).

χ^2 statistics can also be used when sample size is large. If

$$-\left(n-1-\frac{d+g}{2}\right)\ln\left(\Lambda^*\right)>\chi^2_{d(g-1),\alpha}, \tag{4.13}$$

we conclude that the means are different at the significance level α. In (4.13), Λ* is given by (4.12), n is the size of the sample, d is the dimension of the vectors $\mathbf{X}_{ij}$, and g is the number of groups.

Now that we have theoretically described MANOVA it is time for us go back to our ERP example at hand in order to see how we can apply MANOVA on it. The electrodes chosen for the study are F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4. We are interested to study the difference of scalp potentials during odd-ball tasks in the individual electrodes. Each electrode is a group with two observation vectors namely, trial-averaged potential during frequent events and trial-averaged potential during rare events. Note that only MANOVA can tell if the potential associated with two types of events are different in at least one channel.

Now if we are interested to determine at a particular time point whether the difference between the ERP during the odd-ball tasks is significant, we may have to resort to partial lambda statistic. Note that here there are d time points. Since we are interested in studying P300, we would like to see if the difference is significant 300 ms after the stimulus onset. Let 300 ms after the onset is at the time point t (the value of t will depend on the sample frequency). We denote partial lambda by Λ_t^*, defined by

$$\Lambda_t^* = \frac{\Lambda^*\left(X_{ij}^1, X_{ij}^2, X_{ij}^3, \ldots, X_{ij}^d\right)}{\Lambda^*\left(X_{ij}^1, X_{ij}^2, X_{ij}^3, \ldots, X_{ij}^{t-1}, X_{ij}^{t+1}, \ldots, X_{ij}^d\right)}, \tag{4.14}$$

where the numerator and the denominator on the right-hand side are to be calculated by (4.12). $0 \le \Lambda_t^* \le 1$. If

$$\frac{1-\Lambda_t^*}{\Lambda_t^*}\frac{n-g-d+1}{g-1} > F_{g-1,n-g-d+1,\alpha}, \tag{4.15}$$

we conclude that X_{ij}^t contributes significantly for making the difference between ERP recorded during visual odd-ball tasks. Hence the instance at which there is a significant difference in ERP is the tth time point ($1 \le t \le d$).

In the stated experiment we can take the observation vector $\mathbf{X}_{ij}$ as the EEG recorded (and digitized) at the ith channel during the jth task (say for $j=1$ it is the frequent event and for $j=2$ it is the rare event). The entire duration of recording (for each event) is d/Fs second, where Fs is the sample frequency in Hz and n is the total number of trials recorded. Using MANOVA in this way we can not only determine the significant difference between ERP during cognitive tasks but also tell the significant time of occurrence of such a difference.

References

Gibbons, J. D., *Nonparametric Statistical Inference*, McGraw Hill Kogakusha Ltd., Tokyo, Japan, 1971.

Gonzalez, R., Lecture notes #2: Introduction to analysis of variance, University of Michigan, Ann Arbor, MI, 2009, available online at http://www-personal.umich.edu/~gonzo/coursenotes/file2.pdf.

Johnson, R. A. and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed., Pearson Education Inc., Upper Saddle River, NJ, 2002.

Luck, S. J., *An Introduction to Event-Related Potential Technique*, MIT Press, Cambridge, MA, 2005.

Majumdar, K., Constraining minimum norm inverse by phase synchronization and signal power of the scalp EEG channels, *IEEE Trans. Biomed. Eng.*, **56(4)**: 1228–1235, 2009.

Majumdar, K., P. D. Prasad, and S. Verma, Synchronization implies seizure or seizure implies synchronization? *Brain Topogr.*, **27(1)**: 112–122, 2014.

Makeig, S., Auditory event-related dynamics of the EEG spectrum and effects of exposure to tunes, *Eletroencephal. Clin. Neurophysiol.*, **86**: 283–293, 1993.

Murray, M. M., D. Brunet, and C. M. Michel, Topographic ERP analyses: A step by step tutorial review, *Brain Topogr.*, **20**: 249–264, 2008.

5

Binding Problem

Our brain is divided into different functionally specialized areas. Each of them is dedicated to perform certain particular tasks. For example, when person A speaks and another person B listens, the visual cortex of B processes the visual image of A and the auditory cortex of B processes the speech of A. In the end, they both have to be put together to make B understand that the image and sound are coming from the same person. The process of integration of information collected through different sensory areas is an important example of the binding problem. For a nice exposition on the binding problem, see Triesman (1996). Seven different kinds of binding have been mentioned in Triesman (1996).

1. Property binding, in which different properties (e.g., shape, color, and motion) must be bound to the object that they characterize. The 'binding problem' arises here because different aspects of the scene are coded in specialized visual areas, with the size of visual receptive fields increasing at a later level of coding hierarchy.
2. Part binding, in which parts of an object must be segregated from the background and bound together, sometimes across discontinuities, resulting from partial occlusion.
3. Range binding, in which particular values on a dimension (e.g., purple on the color dimension or tilted 20° on orientation dimension, etc.) are signaled by ratios of activity in a few distinct populations of neurons (e.g., those sensitive to blue and to red to signal purple).
4. Hierarchical binding, in which the features of shape-defining boundaries (e.g., orientation, curvature, and closure) are bound to the surface-defining properties that carry them (e.g., luminance, color, texture, motion, and stereoscopic depth).
5. Conditional binding, in which the interpretation of one property (e.g., direction of motion) often depends on another (e.g., depth, occlusion, or transparency).
6. Temporal binding, successive states of the same object must be integrated across temporal intervals, in real and apparent motion and other transformations.
7. Location binding, in which objects are bound to their current locations. Objects and locations appear to be separately coded in ventral and dorsal pathways, respectively, raising what may be the most basic binding problem: linking 'what' to 'where'.

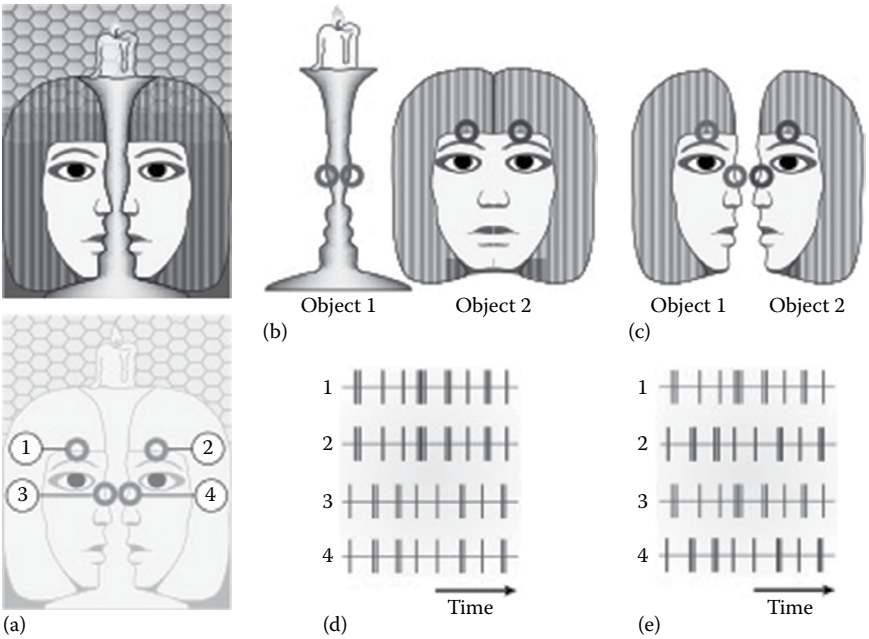


FIGURE 5.1
Temporal binding model: Consider four visual cortical neurons with receptive fields (positions 1–4 in (a)) over image components, the grouping of which changes with transition from one component (say (b)) to another (say (c)). As shown in (d), neurons 1 and 2 should synchronize if the respective contours are part of the one background face and the same should hold for neurons 3 and 4, which represent the contour of a candlestick. When the image is segmented into two opposing faces, the temporal coalition switches to synchrony between 1 and 3 and 2 and 4, respectively, as shown in (e). (Taken from Engel, A.K. et al., *Nat. Rev. Neurosci.*, 2, 704, 2001.)

Most modeling efforts have focused on binding of parts or properties (Triesman, 1996). Figure 5.1 shows a model of temporal binding. Notice that in this model of temporal binding, a synchronous or simultaneous firing of action potentials by two or more neurons indicates the binding. Synchronization is a central concept in brain electrophysiological signal processing, including scalp EEG processing. Synchronization is studied at all scales from single neurons to the whole brain EEG.

5.1 Synchronization

The EEG is processed in order to understand what is happening inside the brain, that is, how the brain is working, albeit very little can be told about the brain from the EEG. Nevertheless, scalp EEG study is the cheapest

and easiest way to study the brain. From the signal processing point of view, any processing of EEG falls under two broad paradigms—studying the rhythm or oscillation and studying the synchronization. Brain oscillations or frequency patterns are studied by time-frequency analysis of the electrophysiological signals (for details, see Van Quyen and Bragin, 2007). The most basic techniques of time-frequency analyses are Fourier transformation and wavelet transformations elaborated in Appendices A and B, respectively.

Since different brain regions can be engaged in the same task, how those regions are transiently oscillating together is an important property to study. This is accomplished by analyzing signals recorded from different brain regions to see how they are evolving together. Different quantitative tools are available for this purpose. In neuroscience literature these tools are broadly known as synchronization measures or coherence measures. The oldest among them is statistical correlation. Let $x(t)$ and $y(t)$ be two digitized signals of same length. Then correlation coefficient r between them is defined by

$$r = \frac{\sum_{i=1}^{N} \left(x(i) - \bar{x}\right)\left(y(i) - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N} \left(x(i) - \bar{x}\right)^2} \sqrt{\sum_{i=1}^{N} \left(y(i) - \bar{y}\right)^2}}, \tag{5.1}$$

where

N is the number of time points present in each signal

$\bar{x}$ denotes the mean of x

From (5.1) it is clear that the correlation coefficient can vary from −1 to 1. The correlation coefficient is measured within a window across both the signals, and then the window is slid along the signals. The value of r is taken at each position of the window and plotted. This plot gives the dynamic correlation between the two signals over a period of time. For multiple signals (say $k \geq 2$ in number) a $k \times k$ matrix is formed, whose i, jth entry is the correlation between ith and jth channels, in a particular window. Let us call this matrix as the *correlation matrix*. Now find all the eigenvalues of this symmetric square matrix and order them. Take only the highest eigenvalue in each window. Plot this highest eigenvalue for all the successive windows (Figure 5.2). This will give the ensemble correlation over time across all the k channels (Schindler et al., 2007).

Although correlation is the oldest and the most popular technique to examine similarity between two or more signals, it is not the most sensitive to small deviations from similarity. For studying the minute changes in transient coupling and decoupling among brain regions, we need a more sensitive measure and this is provided by phase synchronization.

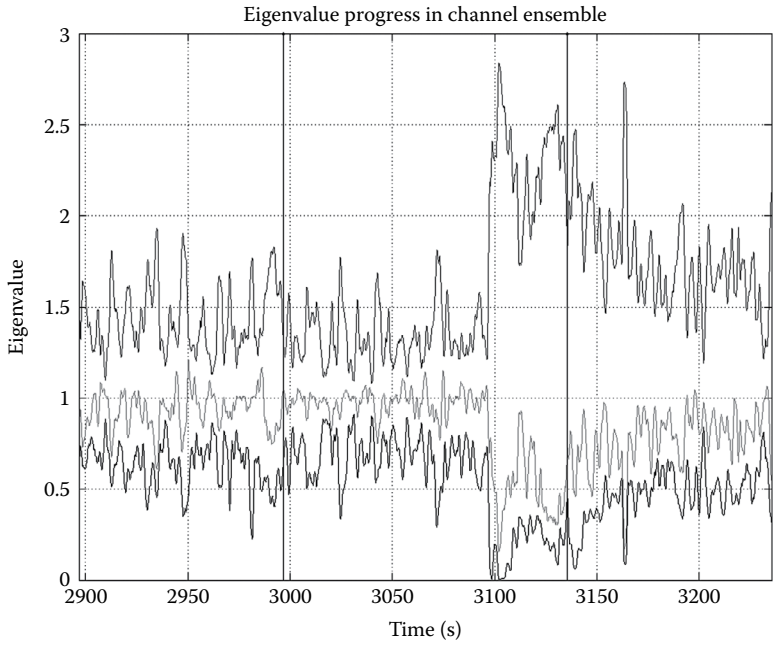


FIGURE 5.2
Plot of eigenvalues of the correlation matrix for three channels over time. The curve at the top is the plot of the highest eigenvalue, which indicates the ensemble correlation among the three channels. Vertical lines indicate seizure onset and offset times.

5.1.1 Phase Synchronization

Consider a simple sinusoidal signal $s(t) = a\sin(2\pi nt + \in)$. The signal $s(t)$ can be changed by changing the amplitude a or frequency n or $\in$. $\in$ is called the *phase* of the signal $s(t)$. Intuitively, two signals have same phase when they have similar curvature at the same time point, as shown in point B of Figure 5.3. Likewise, they have the opposite phase when they have the opposite curvature at the same point, as shown in point A of Figure 5.3. When two sinusoidal signals with same frequency, but with possibly different amplitudes, have a constant difference in phase for a certain duration, the signals are called *phase locked* or *phase synchronous* for that duration. But most of the signals in real life are not sinusoidal signals. In fact, nonperiodic or nonstationary signals (i.e., signals whose statistical properties remain the same over all the windows of a fixed length) cannot completely be decomposed into sinusoidal components. Then how to define phase for those signals?

In Chapter 2 we have dealt with z-transforms, which mapped real-valued signals into the complex plane and helped to model impulse responses with complex functions in a convenient way. This was particularly useful

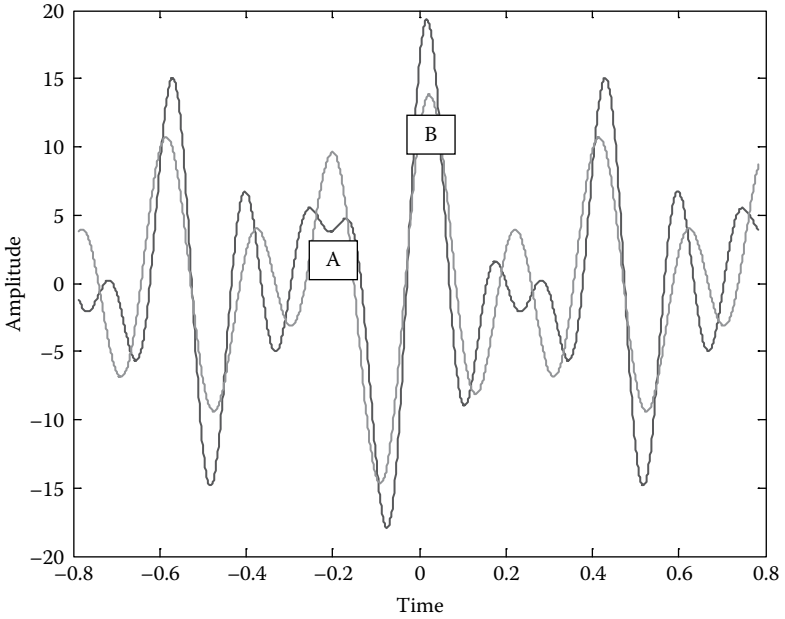


FIGURE 5.3
Two simulated signals. At point A, the two signals are in the opposite phase (phase difference is π). At point B, they are in same phase (phase difference is 0). Since in the entire range the phase difference between the two signals is changing, they are not phase synchronous.

in filter design. Here we will transform the real signals into the complex plane to determine their phase.

Phase is always associated with frequency. So, phase synchronization always means phase synchronization within a frequency band. The narrower the band, the more meaningful is the term 'phase synchronization'. A small ratio between the bandwidth and the mid-band frequency is desirable. For example, phase synchronization between 50 and 60 Hz is more meaningful than phase synchronization between 20 and 30 Hz, because $(60-50)/55 < (30-20)/25$. The most ideal will be the case where phase synchronization is associated with a single integer frequency. It is essential to band-pass filter the signals before measuring the phase synchronization or phase difference between them.

5.1.1.1 Hilbert Transformation Based

The *Hilbert transform* $H(s(t))$ of a signal $s(t)$ is defined as

$$H\left(s\left(t\right)\right) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s\left(\tau\right)}{t-\tau}d\tau. \tag{5.2}$$

The integral may not exist in general, but if its Cauchy principal value (p.v. in short) exists, then that will define the transform. The *Cauchy principal value* is defined as

$$\text{p.v.}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}d\tau=\lim_{\epsilon\to 0}\left[\int_{-\infty}^{t-\epsilon}\frac{s(\tau)}{t-\tau}d\tau+\int_{t+\epsilon}^{\infty}\frac{s(\tau)}{t-\tau}d\tau\right],$$

when the limit exists. So the Hilbert transform can be redefined as

$$H\big(s(t)\big)=\frac{1}{\pi}\text{p.v.}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}d\tau. \tag{5.3}$$

Without loss of generality for any practical purpose, we can define a time domain signal as a real-valued function of nonnegative real variables ($s:\{0\}\cup\Re^{+}\to\Re$, where $\Re$ is the set of real numbers and $\Re^{+}$ is the set of nonnegative real numbers), whose Hilbert transformation exists. How should the Hilbert transformation of a simple function, say $\sin(t)$, look like?

$$\begin{aligned}H\big(\sin(t)\big)&=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin(x)}{t-x}dx=-\frac{1}{\pi}\int_{\infty}^{-\infty}\frac{\sin(t-z)}{z}dz\quad t-x=z,-dx=dz\\&=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin(t)\cos(z)-\cos(t)\sin(z)}{z}dz\\&=\frac{1}{\pi}\sin(t)\int_{-\infty}^{\infty}\frac{\cos(z)}{z}dz-\frac{1}{\pi}\cos(t)\int_{-\infty}^{\infty}\frac{\sin(z)}{z}dz\\&=-\cos(t)\end{aligned} \tag{5.4}$$

because $\int_{-\infty}^{\infty}\frac{\sin(x)}{x}dx=\pi$ and $\int_{-\infty}^{\infty}\frac{\cos(x)}{x}dx=0.$

Similarly, it can be shown that

$$H\big(\cos(t)\big)=\sin(t). \tag{5.5}$$

Equations 5.4 and 5.5 together ensure that the Hilbert transform makes a 90° phase shift (disregard the sign) to trigonometric functions. If the signal is Fourier decomposable (that is, its Fourier series [see Appendix A] converges to the signal at every point), then the Hilbert transformation of the signal is orthogonal to itself, that is, the signal and its Hilbert transform have a 90° phase difference. This motivates us to regard the Hilbert transformation $H(s(t))$ of a signal $s(t)$ to be orthogonal to $s(t)$. (Be aware of caveats like $s(t)=\frac{1}{1+t^2}$, whose Hilbert transform $H\big(s(t)\big)=\frac{t}{1+t^2}$, and therefore $s(t)$ and $H(s(t))$ are not mutually orthogonal at every t.) Let us specifically adopt the following mathematical definition for a signal.

Definition 5.1

A *time domain analog signal* is a function $s:\{0\}\cup\Re^+\to\Re$, such that, (1) its Hilbert transform $H(s(t))$ exists, and (2) $s(t)$ and $H(s(t))$ are mutually orthogonal at each time point t.

A discrete sampled version of $s(t)$ at equal time intervals is called a *digital signal* (please see Chapter 2). Please note that if we go back to the signal $s(t)=a\sin(2\pi nt+\in)$, then in general it will be $s(t)=a(t)\sin(2\pi n(t)t+\in(t))$, where $a:\Re^+\to\Re$, $n:\Re^+\to\Re$ and $\in:\Re^+\to\Re$. This means $s:(\Re^+)^3\to\Re$, and any change in analog $s(t)$ will be captured in the following equation:

$$\frac{ds}{dt}=\frac{\partial s}{\partial a}\frac{da}{dt}+\frac{\partial s}{\partial n}\frac{dn}{dt}+\frac{\partial s}{\partial \in}\frac{d\in}{dt}, \tag{5.6}$$

provided the time domain analog signal is at least piecewise smooth. Equation 5.6 is important from the neural information content of a signal $s(t)$. Whether a superposition of signals of the form $s(t)=a(t)\sin(2\pi n(t)t+\in(t))$ can approximate *any* neural signal is an open question. Even if the answer to this question turns out to be negative, Equation 5.6 does not lose its significance. However, we will explore (5.6) elsewhere, and not in this treatise any further. We will finish this paragraph by observing $s:(\Re^+)^k\to\Re$, $k\geq 1$ in general. Throughout the book we will stick to Definition 5.1. All processings described in this book are performed on digital signals only.

Now we are ready for the analytic form of a time domain signal to be defined on the complex plane. In the complex plane, let us take $s(t)$ along the real axis and $H(s(t))$ along the imaginary axis as shown in Figure 5.4. This is appropriate because they are mutually orthogonal. Now we define the *analytic signal* of $s(t)$ as $s(t)+jH(s(t))$, where $j=\sqrt{-1}$. By *instantaneous phase* (i.e., phase $\phi(t)$ at an arbitrary time point t) of $s(t)$, we mean the phase of the analytic form of $s(t)$, which is defined as

$$\phi(t)=\tan^{-1}\left(\frac{H(s(t))}{s(t)}\right), \tag{5.7}$$

as shown in Figure 5.4.

Let $\phi_1(t)$ and $\phi_2(t)$ be instantaneous phases of the signals $s_1(t)$ and $s_2(t)$, respectively. Then the phase difference between the two signals is $\phi_1(t)-\phi_2(t)$ mod 2π, because the phase difference must lie between $-\pi$ and π or between 0 and 2π. It will be perfectly rational to call the two signals phase synchronous or phase locked over a period $[t_1,t_2]$ if and only if $\phi_1(t)-\phi_2(t)$ mod 2π is constant for $t_1\leq t\leq t_2$. But the practical notion of phase synchronization is slightly more involved, because (1) to demand that $\phi_1(t)-\phi_2(t)$ mod 2π is an exact constant will be too restrictive for most practical purposes, and

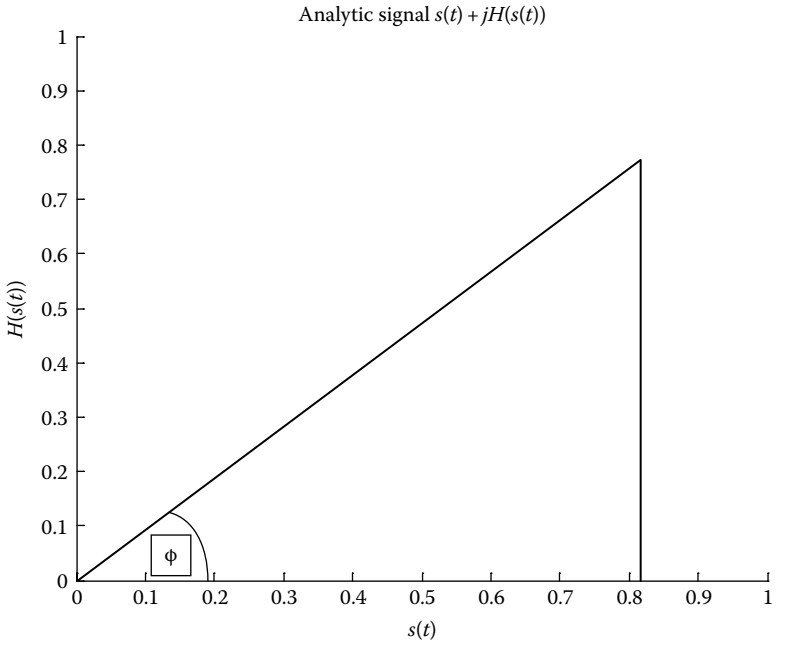


FIGURE 5.4
Analytic signal $s(t) + jH(s(t))$ of a time domain signal $s(t)$, where $j = \sqrt{-1}$, where the instantaneous phase $\phi(t) = \tan^{-1}\left(\dfrac{H(s(t))}{s(t)}\right)$.

(2) whatever phase synchronous or phase locking value will be obtained, what is the guarantee that the value is not due to mere chance? While rectifying (1) is easy, rectifying (2) needs the notion of 'statistical significance testing'. Both of these will have to be taken care of in a useful definition of phase synchronization.

For rectifying (1), let us restate 'phase synchronization' or 'phase locking' from time t_1 to time t_2 as 'approximate phase locking' if the following condition is satisfied for $t_1 \le t \le t_2$

$$\left|\left(\phi_1(t) - \phi_2(t)\right) \bmod 2\pi - C\right| < \delta, \tag{5.8}$$

where C is a value such that $0 \le C \le 2\pi$ and $\delta > 0$, is a 'small' real number. It is clear that in the exact sense of phase synchronization or phase locking from t_1 to t_2, $\delta = 0$ and $\phi_1(t) - \phi_2(t) = C$ for $t_1 \le t \le t_2$. We call 2δ the *bin size*. Why? That will become clear soon.

For rectifying (2), we need to show that the phase synchronization value for the signal pair at hand is higher than that on a collection of random signal pairs. For this, first we will have to generate a reasonable-sized sample of

random signal pairs. This can be done by taking Gaussian noise pairs or by generating random signals out of a real signal using the 'shifted surrogate' method. The second one is more appealing because it preserves the amplitude of the original signals. Only values at the time points are randomly permuted or shuffled among themselves exactly like the shuffling of a pack of cards. This can easily be done in MATLAB® by using the 'randperm' function. If there are N time points in a signal, it is possible to generate at most $N!-1$ different random signals out of that signal. We only need 100 to 1000 pairs out of them to make the sample. This is, of course, not foolproof, but this works well and is accepted in the neural signal–processing community. Now, for each pair of actual signals as well as random signals we perform the following.

We divide the interval $[0,\ 2\pi]$ into, say k number of bins. We have to choose a fixed window size, say with N_w time points, for the signals. For a pair of signals, for each time point t in a window, there will be a value $\phi_1(t)-\phi_2(t) \bmod 2\pi$, which should be placed in the ith interval if $\frac{i}{k}2\pi \le \phi_1(t)-\phi_2(t) \bmod 2\pi < \frac{i+1}{k}2\pi$. After all the N_w such values are distributed among the k bins, the frequency distribution of values in a bin is calculated by dividing the number of values in the bin by N_w. Let the frequency at the ith bin be p_i. Clearly, p_i is a probability measure. So we are in a position to calculate the Shannon entropy S of the distribution of phase-locking values across the entire signal window, which is $S=-\sum_{i=1}^{k} p_i \log p_i$. This entropy is maximum (for the discrete case) when p_i is a uniform probability density measure (for the continuous case, it will be the Gaussian density function). So $S_{\max}=\log k$. Let us define γ such that,

$$\gamma = \frac{S_{\max}-S}{S_{\max}}. \tag{5.9}$$

Clearly $0 \le \gamma \le 1$. γ is called the *phase synchronization* measure between two signals within the chosen window. Now the window can be slid to have phase synchronization values between the signal pairs across the entire time duration. Note that the bin size is $2\pi/k$, which is the allowable range for the phase difference value to vary about a central tendency C. Therefore, $(2\pi/k)=2\delta$ or $\delta=\pi/k$.

If two signals are phase locked or phase synchronous over a certain duration, then the phase difference values in that duration should lie in a single bin or only in a few contiguous bins. In that case, S will have a smaller value and γ in Equation 5.9 will be closer to 1. On the other hand, if the two signals are not phase synchronous, the phase difference values across the signals will be distributed more uniformly among all the bins and S will be closer to $S_{\max}$ and γ will have a small value.

Now establishing the statistical significance of the synchronization values between the two signals at hand is straightforward. Calculate phase synchronization value using (5.9) across all pairs of surrogate signals as well

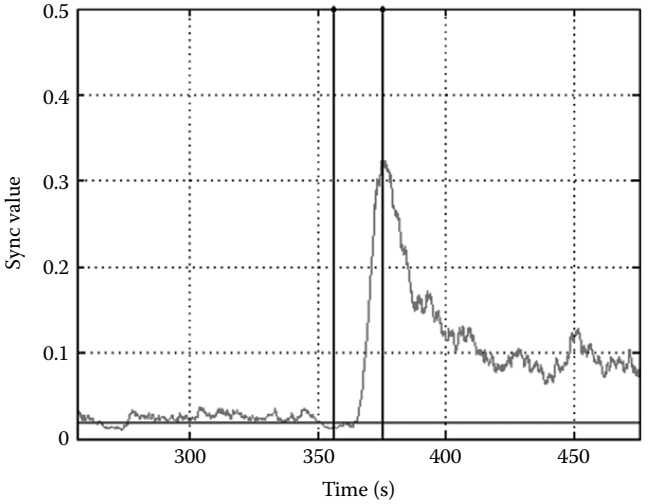


FIGURE 5.5
Hilbert phase synchronization between a pair of depth EEG signals from a patient with epilepsy. Here continuous sliding window has been used. The horizontal line denotes the statistical significance value, that is, any synchronization value above the red line is statistically significant. Vertical lines indicate seizure onset and offset points. (Adapted from Prasad, P.D. et al., *Clin. EEG Neurosci.*, 44(1), 16, 2013.)

as the pair of real signals at hand for processing by sliding the window in a suitable manner (continuous sliding, that is sliding the window by one point at a time, is quite popular). Take the average synchronization value across the whole signal for the surrogate pairs (the maximum can also be taken in place of average). Sort them in descending order. Take the value that is greater than or equal to, say 95% of the values. If the phase synchronization value (calculated using (5.9)) at each point (the value for the sliding window is the value for a point in the signal, which is particularly convenient for a continuously sliding window) is greater than the value taken to be the 95 percentile value across all surrogate signal pairs (Figure 5.5), then it is accepted as 'statistically significant' with a probability of error less than or equal to 0.05 (Figure 5.5).

It is important to note that (5.9) is not the only way to define a Hilbert transform–based phase synchronization measure. For example the quantity $\frac{1}{N}\left|\sum_{i=1}^{N}\exp\left(j\left(\phi_1(t)-\phi_2(t)\right)\right)\right|$ (we will use it in the next section), where N is the number of time points, is also a popular measure of Hilbert transform–based phase locking (popularly known as the *phase locking value*). In neuroscience, it is important for the signals from different brain regions to become synchronous for a short period of time and then again become asynchronous. In fact, transitions from the asynchronous to the synchronous state and then again from the synchronous to the

asynchronous state are essential for the performance of various tasks by the brain. Abnormalities in these cycles have been observed during epileptic seizures and schizophrenic hallucinations. Therefore, having an estimate of phase synchronization like that in (5.9) is important for studying the brain, which gives an unbiased normalized measure both during high synchronization and high asynchronization and also for states in between these two. For a nice review on Hilbert transformation–based phase synchronization, see Rosenblum et al. (2001).

5.1.1.2 Wavelet Transformation Based

A signal may be considered to be a superposition of different waveforms. Then a signal can also be decomposed into those waveforms, provided they can be identified in a practical way. Wavelets are mathematical models of those waveforms (see Appendix B). For wavelet analysis of a signal, first a suitable family of wavelets (there is a generic form for each family, which is called the *mother wavelet*, and then varying the parameter values in the mother wavelet other wavelets of the family are derived, which are called *daughter wavelets*) is chosen (see Appendix B for an elaboration). Next, an appropriate window is chosen. Then suitable members of the wavelet family are convoluted with the signal within the window and the window is slid as usual. The convolution is expected to extract waveforms from the signal containing features according to the characteristics of the wavelet chosen (unfortunately there is no rigorous theory yet about how to choose a wavelet; it is largely subjective).

Phase is, of course, a very important feature or characteristic of a signal, which is strongly associated with frequency and changes with respect to time in the same signal. It is therefore quite legitimate to expect that some kind of a time frequency analysis would be able to identify the phase of a signal at each time and frequency. The wavelet of choice for extracting the phase from a signal is the Morlet (or Gabor) wavelet (Figure 5.6).

The Morlet wavelet is mathematically represented by

$$\psi_{t,f}(x) = \sqrt{f}\exp\left(j2\pi f(x-t)\right)\exp\left(-\frac{(x-t)^2}{2\sigma^2}\right). \tag{5.10}$$

This is the product of a sinusoidal wave at a frequency f, with a Gaussian function centered at time t and a standard deviation σ, which is inversely proportional to f. Figure 5.6 shows its imaginary part for $t=0$, $f=5$, and $\sigma=1$. Now the Morlet (or Gabor) wavelet (5.10) will have to be convoluted with each signal to extract its phase for frequency f.

$$W_s(t,f) = \int_{-\infty}^{\infty} s(x)\psi^*_{t,f}(x)\,dx, \tag{5.11}$$

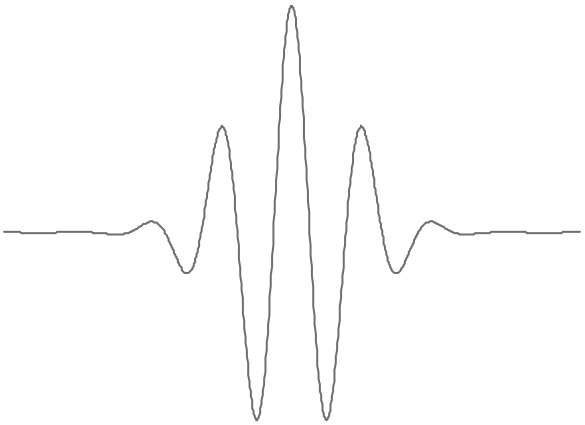


FIGURE 5.6
Morlet wavelet. A convolution with it accentuates the peak and trough parts of the signal, when the signal peak is superimposed on the wavelet peak during the sliding of the window. The same is true for troughs.

which will give the wavelet coefficients (just like the nth Fourier coefficient is obtained by convoluting the signal with the conjugate of $\exp(j2\pi nx)$) in terms of t and f. Here ψ^* is the conjugate of ψ. The wavelet coefficient given by (5.11) is supposed to contain phase information of the frequency band $\left[f-\frac{4f}{6f\sigma}, f+\frac{4f}{6f\sigma}\right]=\left[f-\frac{2}{3\sigma}, f+\frac{2}{3\sigma}\right]$ as described in Van Quyen et al. (2001). Next, the phase difference is defined by the formula

$$\exp\left(j\left(\phi_{s1}(f,t)-\phi_{s2}(f,t)\right)\right)=\frac{W_{s1}(t,f)W_{s2}^{*}(t,f)}{\left|W_{s1}(t,f)\right|\left|W_{s2}^{*}(t,f)\right|}. \tag{5.12}$$

By the *phase locking value* (PLV) between the signals s_1 and s_2, within a time interval $[a,b]$, we mean the quantity

$$\Gamma=\frac{1}{b-a}\int_{a}^{b}\exp\left(j\left(\phi_{s1}(f,t)-\phi_{s2}(f,t)\right)\right)dt, \tag{5.13}$$

which is calculated with the help of Equation 5.12. The statistical significance is determined exactly as described in the previous section.

There is no mathematical proof that (5.9) and (5.13) will always give identical results. But in a large number of cases they have given equivalent outcomes (see Van Quyen et al. (2001) for a data-driven equivalence between the two phase difference measures). The Hilbert transform–based phase synchronization measure in Equation 5.9 is based on the notion of instantaneous phase, which is more precise than the ratio of the

products of wavelet coefficients containing phase-difference information, as presented in Equation 5.12.

5.1.1.3 Fourier Transformation Based

The idea of phase synchronization is about four centuries old. It was first described by Christian Huygens for two pendulums oscillating simultaneously. The notion of phase synchronization that was conceived was essentially phase synchronization between two sinusoidal signals. Fourier transformation–based phase synchronization captures the geometric flavor of that classical notion. Also, unlike the previous two measures, the Fourier transform–based phase synchronization measure works in a frequency domain rather than a time domain.

Let $x_j(t)$ and $x_k(t)$ be the EEG signals collected from the jth and kth electrodes, respectively. Their Fourier expansions can be written as follows (Majumdar, 2009):

$$x_j(t) = \frac{a_{j0}}{2} + \sum_{n=1}^{\infty} \left(a_{jn}^2 + b_{jn}^2\right)^{1/2} \sin\left(\frac{2\pi nt}{p} + \alpha_{jn}\right) \tag{5.14}$$

with phase α_{jn} given by $\alpha_{jn} = \tan^{-1}\left(\frac{a_{jn}}{b_{jn}}\right)$, and

$$x_k(t) = \frac{a_{k0}}{2} + \sum_{n=1}^{\infty} \left(a_{kn}^2 + b_{kn}^2\right)^{1/2} \sin\left(\frac{2\pi nt}{p} + \alpha_{kn}\right) \tag{5.15}$$

with phase α_{kn} given by $\alpha_{kn} = \tan^{-1}\left(\frac{a_{kn}}{b_{kn}}\right)$.

a_{jn}, b_{kn} are determined by FFT within a suitably chosen time window. p is the duration of the signal (no real signal can remain alive for ever; it will die down after a time lapse).

If two signals are phase synchronous over a duration of time, the difference between the pair's instantaneous phases should remain close to a constant value over that duration. In other words, the phase difference at the time instance t and that of the next time instance $t+1$, when both are within that duration, will be very close to each other. We can mathematically express it by $mean(E(t)) \approx 0$ and $std(E(t)) \approx 0$, where *std* stands for standard deviation and $E(t)$ is defined as $E(t) = |(\phi_1(t) - \phi_2(t)) - (\phi_1(t+1) - \phi_2(t+1))|$. The mean and standard deviation are taken across a window and then the window is slid across the signal pair. A possible normalized phase synchronization measure is then $\frac{1}{1 + mean\left(E(t)\right)}$. It is interesting to note that this measure gives quite close results to that of (5.9) when the signal pairs are not band-pass filtered

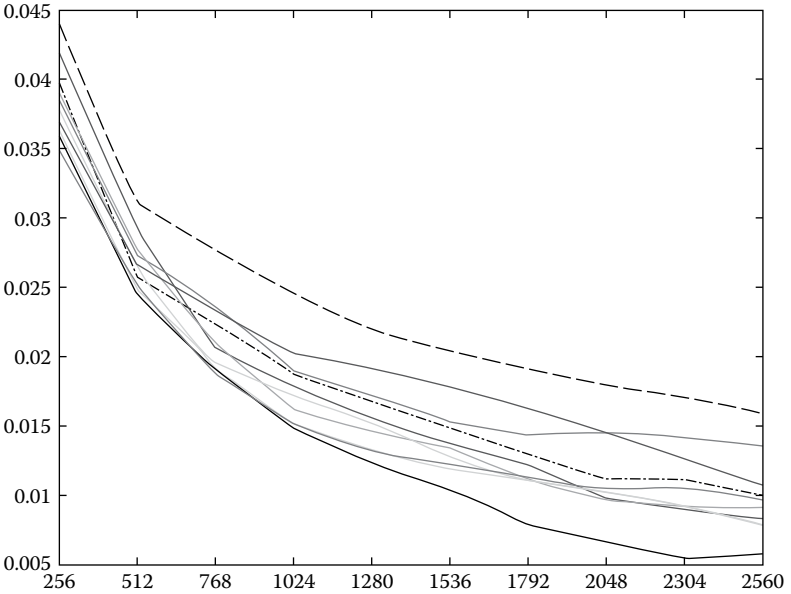


FIGURE 5.7
Difference of Hilbert transform–based phase difference measure by Shannon entropy and $1/(1+mean(E(t))$ (along the ordinate) with respect to window length in a number of time points (along the abscissa) for ten pairs of shifted surrogate signals.

and window size is 'large' (Figure 5.7). The mismatch between the two measures becomes large if a narrow bandwidth is chosen.

The same idea can be carried over to the frequency domain. By carrying out fast Fourier transform (FFT) of the signals we can obtain the phase associated with each frequency component with the help of the corresponding Fourier coefficients, as shown in (5.14) and (5.15). If $x_j(t)$ and $x_k(t)$ are phase synchronous signals with almost zero phase lag across all the harmonics, we call them *tightly coupled signals* (TCS). This situation occurs, for example, when a signal from a strong source is received through multiple sensors. All the sensors will then be pair-wise TCS. This happens during the EEG source localization and may give us an additional handle to better localize the cortical sources of scalp EEG (Majumdar, 2009). In a tightly coupled EEG signal pair $x_j(t)$ and $x_k(t)$, the following will hold

$$\alpha_{j1} - \alpha_{k1} \approx \cdots \approx \alpha_{jn} - \alpha_{kn} \approx \cdots, \tag{5.16}$$

which implies

$$\frac{a_{j1}b_{k1} - a_{k1}b_{j1}}{a_{j1}a_{k1} + b_{j1}b_{k1}} \approx \cdots \approx \frac{a_{jn}b_{kn} - a_{kn}b_{jn}}{a_{jn}a_{kn} + b_{jn}b_{kn}} \approx \cdots \tag{5.17}$$

This happens because $\tan^{-1}x \approx x$ as $x \to 0$. Equation 5.17 implies that both the mean and the standard deviation of $E(n)$, which is defined as

$$E(n) = \left| \frac{a_{jn}b_{kn} - a_{kn}b_{jn}}{a_{jn}a_{kn} + b_{jn}b_{kn}} - \frac{a_{jn+1}b_{kn+1} - a_{kn+1}b_{jn+1}}{a_{jn+1}a_{kn+1} + b_{jn+1}b_{kn+1}} \right|, \tag{5.18}$$

be small. Since we are interested in only tight coupling, we have taken $E(n)$ for all n irrespective of the power associated with the band n (in a real signal, every band has a power associated with it, no matter however small). The measure of tight coupling between $x_j(t)$ and $x_k(t)$ is given by the *syn* function defined as

$$syn\left(x_j(t), x_k(t)\right) = syn(j,k) = \frac{1}{1 + mean\left(E(n)\right)}. \tag{5.19}$$

Equation 5.19 gives a measure of (very tight) phase synchronization between $x_j(t)$ and $x_k(t)$. The *syn* function can also be defined as $syn\left(x_j(t), x_k(t)\right) = syn(j,k) = \frac{1}{1 + mean\left(E(n)\right) + std\left(E(n)\right)}$ (Majumdar, 2009), but the measure is more jittery due to the term $std(E(n))$ in the denominator. It is obvious that the time complexity of the *syn* function calculation algorithm equals that of the FFT.

One major concern about (5.19) is that it is a nonlinear measure, which is an impediment for this measure to compare favorably with the Hilbert transform–based measure in (5.9). Equation 5.19 is reliable for high *syn* values, which means signals are highly phase synchronous. But low *syn* values do not necessarily mean that the signals are highly phase asynchronous. Note that $mean(E(n)) \leq 2\pi$ and therefore *syn* values never becomes 0. Of course, it can be made to take the 0 value by subtracting a suitable constant term.

5.1.2 Other Synchronizations

The three phase synchronization measures described in the last section are all linear measures, in the sense that Hilbert, wavelet, and Fourier transforms are all linear transforms (although (5.9), (5.13), and (5.19) are all nonlinear formulae). In this section we will discuss nonlinear measures of synchronization. The foremost of them is perhaps mutual information.

$$I(x_k, x_l) = H(x_k) - H(x_k | x_l) = \sum_{x_k, x_l} p(x_k, x_l) \log \frac{p(x_k, x_l)}{p(x_k)p(x_l)}, \tag{5.20}$$

where $I(x_k, x_l)$ is *mutual information* between channels x_k and x_l, which is a measure of dependence between the two signals. $H(x)$ is the Shannon entropy of

signal x and $H(x \mid y)$ is the Shannon entropy of x while the Shannon entropy of y is already given. $p(x,y)$ is the joint probability density of x and y (Cover and Thomas, 2006).

Close to mutual information is an embedding dimension–based synchronization measure, called the *synchronization likelihood* (Stam and van Dijk, 2002). This is actually a multivariate synchronization measure among two or more EEG channels. Let us take M EEG channels $x_{k,i}$, where $k \in \{1, \ldots, M\}$ and $i \in \{1, \ldots, N\}$, N is a chosen number of time points on the signal $x_{k,.}$. From each of the M time series, embedded vectors $X_{k,i}$ are reconstructed with time delay embedding

$$X_{k,i} = \left(x_{k,i}, x_{k,i+l}, x_{k,i+2l}, \ldots, x_{k,i+(m-1)l}\right), \tag{5.21}$$

where

l is the *lag*

m is the *embedding dimension*

For each time series $x_{k,.}$ and each time point i we define a probability measure $p^{\in}_{k,i,j}$, which is the probability that embedded vectors $X_{k,i}$ and $X_{k,j}$ are close to each other by an amount smaller than $\in > 0$.

$$p^{\in}_{k,i,j} = \frac{1}{2(w_2 - w_1)} \sum_{\substack{j=1 \\ w_1 < |i-j| < w_2}}^{N} \theta\left(\in - \left|X_{k,i} - X_{k,j}\right|\right), \tag{5.22}$$

where θ is a Heaviside function given by $\theta(x) = 0$ for $x \leq 0$ and $\theta(x) = 1$ for $x > 0$. Here w_1 is a window at least as large as the autocorrelation time, and w_2 is another window that sharpens the time resolution of the synchronization measure chosen to obey the inequality $w_1 \ll w_2 \ll N$ (Stam and van Dijk, 2002).

Now a value $p_{ref} \ll 1$ is to be chosen so that $p^{\in_{k,i}}_{k,i,j} = p_{ref}$. For each k and i, the *critical distance* $\in_{k,i}$ is to be chosen so that the equality is satisfied for all j according to (5.22). We actually consider all pairs i,j, such that $w_1 < |i-j| < w_2$, for which the number of channels $H_{i,j}$ have pair-wise signal vectors $X_{k,i}$ and $X_{k,j}$ that are close to each other by an amount of $\in_{k,i}$ or less. Clearly, $H_{i,j}$ is given by

$$H_{i,j} = \sum_{k=1}^{M} \theta\left(\in_{k,i} - \left|X_{k,i} - X_{k,j}\right|\right). \tag{5.23}$$

Clearly, $H_{i,j} \in \{0, \ldots, M\}$, and gives the number of synchronous signals. Let us define the *synchronization likelihood* $S_{k,i,j}$ for each channel k and each pair i,j as

$$\text{if } \left|X_{k,i} - X_{k,j}\right| < \in_{k,i} \quad \text{then } S_{k,i,j} = \frac{H_{i,j} - 1}{M - 1} \tag{5.24}$$

$$\text{if } \left|X_{k,i} - X_{k,j}\right| \geq \epsilon_{k,i} \quad \text{then } S_{k,i,j} = 0. \tag{5.25}$$

By averaging over all j we finally obtain the synchronization likelihood,

$$S_{k,i} = \frac{1}{2(w_2 - w_1)} \sum_{w_1 < |j-i| < w_2}^{N} S_{k,i,j}. \tag{5.26}$$

The synchronization likelihood is a measure that describes how strongly channel k at time i is synchronized to all the other $M-1$ channels. The synchronization likelihood takes on values between p_{ref} and 1. $S_{k,i}=p_{ref}$ corresponds to the case where all M time series are uncorrelated, and $S_{k,i}=1$ corresponds with maximal synchronization of all M time series. The value p_{ref} can be set at an arbitrarily low level and does not depend on the properties of the time series, nor is it influenced by the embedding parameters.

5.1.3 Multivariate Analysis

An approach for extending synchronization between a pair of signals to more than two signals has been described at the beginning of this chapter (e.g., the highest eigenvalue plot in Figure 5.2). Also, the synchronization likelihood described in Section 5.1.2 is a synchronization measure suitable for more than two channels. In this section, we will discuss about some more synchronization measures for more than two channels.

Recently a quantitative measure to define an ordering of the synchronization matrices based on Frobenius norm of a matrix has been proposed (Prasad et al., 2014). If synchronization measure between ith and jth channels a_{ij} (say, by the Hilbert transformation–based method) is the ijth entry of a matrix A (a plot of A for two hemispheres with color-coded a_{ij} has been presented in Figure 5.8), then the *entry-wise norm* $\|A\|_p$ is defined in Equation 5.27, where p is a positive number and n is the number of channels in the ensemble. *Frobenius norm* $\|A\|_F$ is chosen by substituting $p=2$ in Equation 5.27.

$$\|A\|_p = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \left(a_{ij}\right)^p \right)^{1/p}. \tag{5.27}$$

$$\gamma = \sqrt{\frac{\|A\|_F^2 - n}{n^2 - n}}. \tag{5.28}$$

The ensemble measure is computed on a 0–1 scale as given in Equation 5.28, where n is the number of channels in the cluster (Figure 5.9). Here, a value of 0

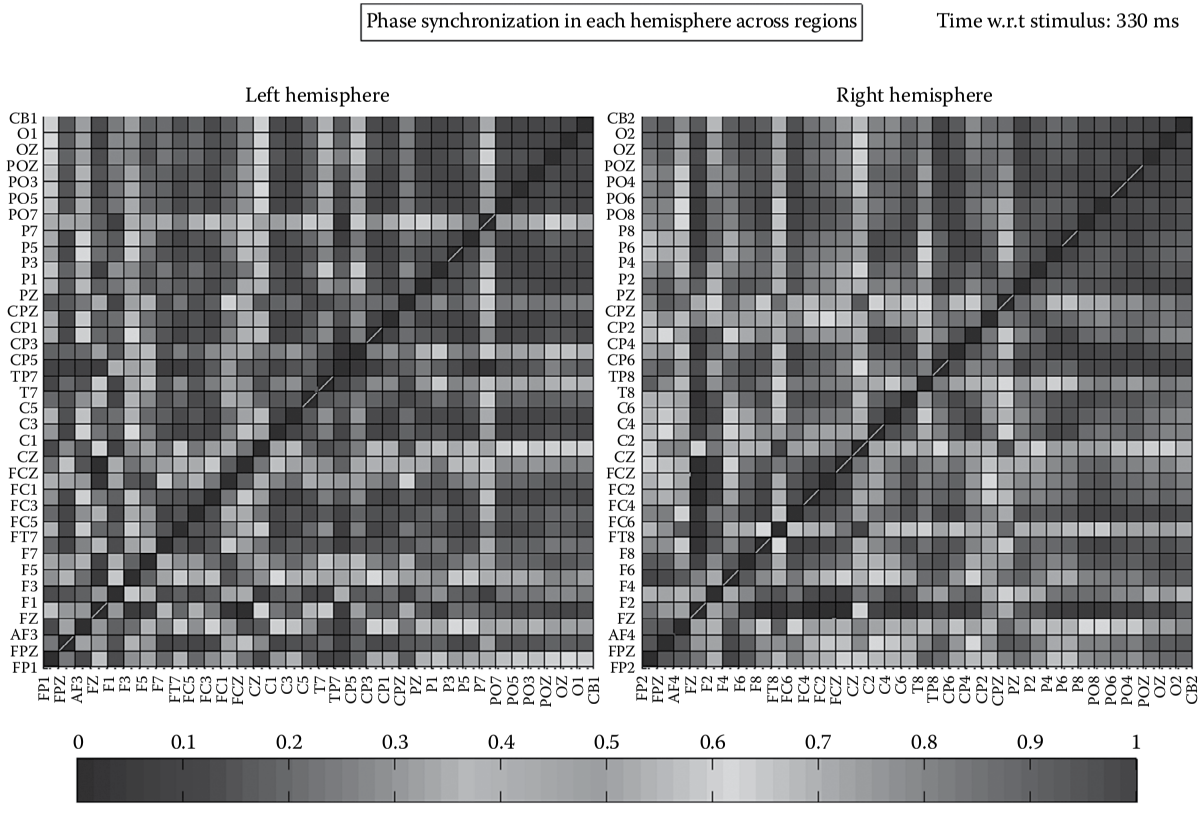


FIGURE 5.8
Plot of pairwise Hilbert phase synchronization for the two hemispheres 330 ms after an audiovisual stimulus onset. Values in the diagonal entries have been kept zero (actually they are uniformly one) for better visualization.

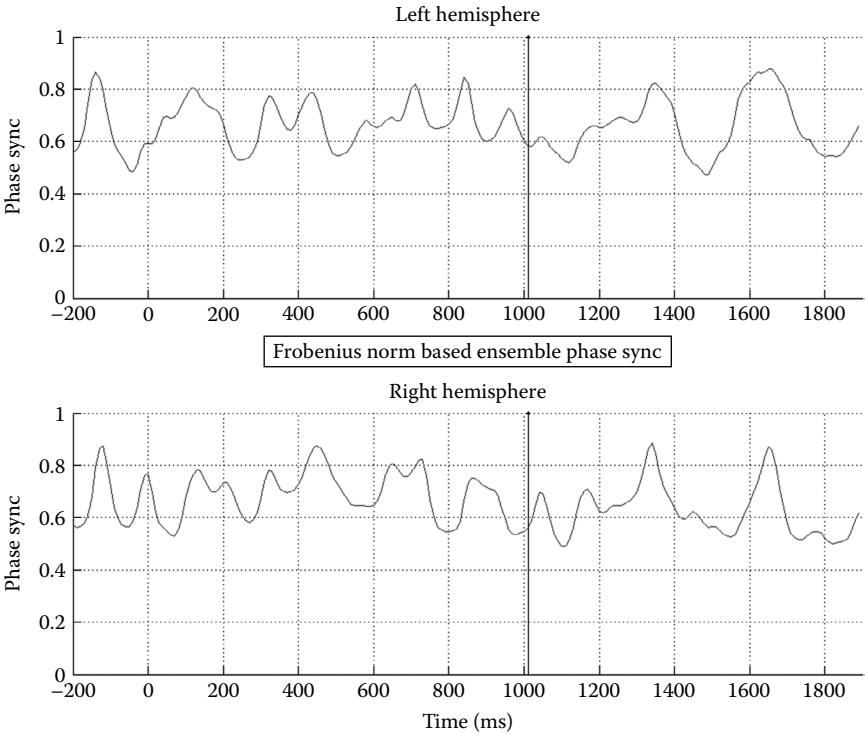


FIGURE 5.9
Ensemble hemispheric synchronization of EEG averaged across all trials for the same subject whose one synchronization snapshot for a short duration within a trial has been presented in Figure 5.8. The vertical line indicates average response latency across all the trials by the subject.

corresponds to the lowest or no-ensemble synchronization while, 1 corresponds to the highest. This enables comparison of different clusters using the proposed ensemble measure, and thus one can tell which ensemble or which subject exhibits greater synchronization.

There is another recent notion of multichannel synchronization called *peak synchronization* (Biswas et al., 2014). It can be mathematically proved that in any one-dimensional time domain digital signal peak and trough are the most significant occurrences (Majumdar and Jayachandran, 2016). In the case of EEG, ECoG, and LFP, peaks (spikes) signify heightened synchronous neural activities. When peaks occur in one or more signals, in the latter case almost at the same time, it is called an *event* (Wilson and Emerson, 2002. Notice the difference from 'event' in an event-related potential (ERP), where an *event* is a sensory stimulus.). Successful event detection in multidimensional neural signals will involve (1) detection of peaks (spikes) and (2) determining if peaks across different signals are occurring simultaneously (with zero or positive time lag). We described a peak detection algorithm in

Section 4.2.1. Here we will describe an algorithm that determines whether peaks in two or more different signals are occurring simultaneously.

Let p_k be a sequence of signals, where $p_k(t)$ is the converted tth time point of the kth signal S_k. The conversion of time points of the original signals happens in the following manner:

$$\begin{aligned} p_k(t) &= 1 \quad \text{if } S_k(t) \text{ is a peak,} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

We first consider two signals and then generalize the measure to more than two signals. Let us define the following quantities, whose significance will become clear subsequently.

$$\begin{aligned} f_k(t) &= \sum_{j=-n}^{n} a_j p_k(t+j) \\ I_k(t) &= 1, \quad \text{if } p_k(t) = 0 \\ &= 0, \quad \text{if } p_k(t) = 1 \end{aligned} \tag{5.29}$$

$k \in \{1, 2\}$, a_j are the weights, obtained by appropriately segmenting a probability density function into strips as shown in Figure 5.10. For finding the

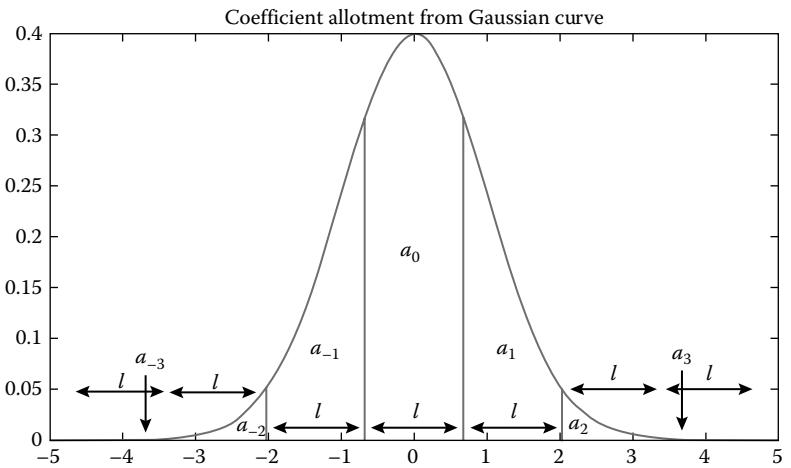


FIGURE 5.10
In the above figure, the length l is determined from the value assigned to a_0 (0.5 in the above figure). The other weights are determined by the area of adjacent strips of length l. The tail area rapidly decreases, the further we move from the origin. Here, the tail area becomes insignificant ($<10^{-4}$) beyond ±4.7, and hence strips after ±4.7 are ignored. This determines the coefficient vector, which in this case is $(a_{-2}, \ldots, a_0, \ldots, a_2)$. (Taken from Biswas, R. et al., *IEEE Trans. Signal Proc.*, 62(17), 4391, 2014.)

peak synchrony, one needs to take a symmetric probability density function, nondecreasing before 0 and nonincreasing afterwards. Here, let $2n + 1$ be the number of strips having area not negligibly small (Figure 5.10). This number is odd, because central strip + places for n peaks on either side of it.

The measure of peak synchronization between S_1 and S_2 at the tth time point, $\phi_{1,2}(t)$, is given by

$$\begin{aligned}\phi_{1,2}(t) &= I_1(t) f_1(t) p_2(t) + I_2(t) f_2(t) p_1(t) \\ &= \left(\sum_{k=1}^{2} (f_k I_k)(t)\right)\left(\sum_{k=1}^{2} p_k(t) - \left(\sum_{k=1}^{2} (f_k I_k p_k)(t)\right)\right)\end{aligned} \tag{5.30}$$

For most practical purposes, one can use the normal density function with mean 0 and scale (standard deviation) 1 as the weight determination function. A mathematical proof of this assertion has been presented in Biswas et al. (2014).

The intuitive idea behind the peak synchronization measure is that given any peak in S_1 at time point t we look for a peak in S_2. If $S_2(t)$ is a peak, we assign the highest weight (a_0) to the peak $S_2(t)$. The peaks on either side of the peak $S_2(t)$ are assigned weights a_j's (other than a_0). Weights are decreasing in magnitude with increasing distance of the peak from $S_2(t)$. The weights are obtained by taking areas of strips, of equal width, under the probability density function. The width of the strips is determined by taking the area of the central strip equal to a_0 (Figure 5.10). $f_k(t)$ captures the extent of the presence of peaks around the time point t. The role of $I_k(t)$ is to average when both $S_1(t)$ and $S_2(t)$ are peaks and $\phi_{1,2}(t)$ is the sum of measure of synchronization of peaks around $S_2(t)$ with the peak $S_1(t)$ and that of peaks around $S_1(t)$ with the peak $S_2(t)$.

It is easy to see that generalization of (5.30) to more than two signals will lead to

$$\begin{aligned}\phi_{1,2,\ldots,r}(t) &= \frac{1}{\binom{r}{2}}\left[\left(\sum_{k=1}^{r} (f_k I_k)(t)\right)\left(\sum_{k=1}^{r} p_k(t)\right) - \left(\sum_{k=1}^{r} (f_k I_k p_k)(t)\right)\right] \\ &= \frac{1}{\binom{r}{2}} \sum_{i=1}^{r-1} \sum_{j=i+1}^{r} \phi_{i,j}(t),\end{aligned} \tag{5.31}$$

where $\phi_{1,2,\ldots,r}(t)$ denotes the ensemble synchronization across the cluster of all the r ($r \geq 2$) channels. A mathematical proof of (5.31) appears in Biswas et al. (2014). In other words, generalization of the pairwise peak synchronization measure to the multichannel case, say to a cluster of r many signals

for $r \geq 2$, would merely be average pairwise synchronization over all the signal pairs in the cluster. The factor $1/(r/2)$ ensures the comparability of values obtained for clusters of signals of different sizes.

Pairwise synchronization (5.30) takes $O(N)$ time to compute, where N is the number of time points in the signals. Generalization to r signals ($r \geq 2$), as given by (5.31), takes $O(r^2N)$ time to compute (Biswas et al., 2014). Since r is fixed and usually $N \gg r$, even the multichannel peak synchronization measure calculation algorithm can be considered to be linear time executable for all practical purposes and therefore suitable for online implementation.

Peak detection (discussed in Chapter 4) and peak synchronization together across multiple channels will help detect events in the central nervous systems. One important application can be in epilepsy research. In simultaneous EEG-fMRI recording of epilepsy patients, interictal spiking activities (in the EEG) and their hemodynamic correlates (in the fMRI) are usually studied. If interictal spikes appear in different channels with a slight time lag, the peak synchronization measure will be able to detect the event. The hemodynamic response of the entire event can be studied rather than that of individual spikes.

Another potential application of peak synchronization may be in measuring synchronization between two signals at different frequencies, where phase synchronization measure will be meaningless (Figure 5.11). This can have many potential applications in neuroscience.

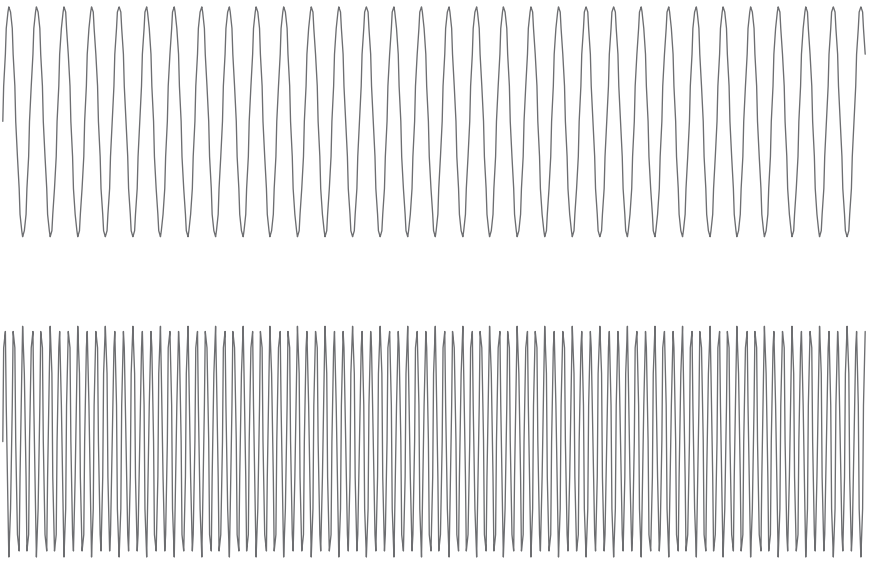


FIGURE 5.11
Two sinusoidal signals with different frequencies. Here phase synchronization between the signals does not make sense, but peak synchronization is well defined.

References

Biswas, R., K. Khamaru, and K. K. Majumdar, A peak synchronization measure for multiple signals, *IEEE Trans. Signal Proc.*, **62(17)**: 4391–4399, 2014.

Cover, T. M. and J. A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, Hoboken, NJ, 2006.

Engel, A. K., P. Fries, and W. Singer, Dynamic predictions: Oscillation and synchrony in top-down processing, *Nat. Rev. Neurosci.*, **2**: 704–716, 2001.

Majumdar, K., Constraining minimum norm inverse by phase synchronization and signal power of the scalp EEG channels, *IEEE Trans. Biomed. Eng.*, **51(4)**: 1228–1235, 2009.

Majumdar, K. and S. Jayachandran, A geometric analysis of time series leading to information encoding and a new entropy measure, *J. Comput. Appl. Math.*, 2017 (in press).

Prasad, P. D., H. N. Halahalli, J. P. John, and K. Majumdar, Single-trial EEG classification using logistic regression based on ensemble synchronization, *IEEE J. Biomed. Health Inform.*, **18(3)**: 1074–1080, 2014.

Prasad, P. D., S. Vishakha Dutta, and K. K. Majumdar, Enhanced phase and amplitude synchronization towards focal seizure offset, *Clin. EEG Neurosci.*, **44(1)**: 16–24, 2013.

Rosenblum, M., A. Pikovsky, C. Schafer, P. A. Tass, and J. Kurths, Phase synchronization: From theory to data analysis, in: F. Moss and S. Gielen (eds.), *Neuro-Informatics and Neural Modeling*, pp. 279–321, Elsevier, Amsterdam, the Netherlands, 2001.

Schindler, K., H. Leung, C. E. Elger, and K. Lehnertz, Assessing seizure dynamics by analyzing the correlation structure of multichannel intracranial EEG, *Brain*, **130(1)**: 65–77, 2007.

Stam, C. J. and B. W. van Dijk, Synchronization likelihood: An unbiased measure of generalized synchronization in multivariate data, *Physica D*, **163**: 236–251, 2002.

Triesman, A., The binding problem, *Curr. Opn. Neurobiol.*, **6**: 171–178, 1996.

Van Quyen, M. L. and A. Bragin, Analysis of dynamic brain oscillations: Methodological advances, *Trends Neurosci.*, **30(7)**: 365–373, 2007.

Van Quyen, M. L., J. Foucher, J.-P. Lachaux, E. Rodriguez, A. Lutz, J. Martinerie, and F. J. Varela, Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony, *J. Neurosci. Methods*, **111**: 83–98, 2001.

Wilson, S. B. and R. Emerson, Spike detection: A review and comparison of algorithms, *Clin. Neurophysiol.*, **113**: 1873–1881, 2002.

6

Epilepsy Research

Currently, epilepsy is ranked the fourth most common neurological disorder after (cerebral) stroke, Alzheimer's, and migraines (a few years earlier it used to be ranked second only after stroke). Worldwide, about 1% of the population is estimated to be affected by epilepsy. It is estimated that the overall prevalence of epilepsy is 5.9–10 per 1000 in India (Rana, 2015). One of the most important applications of EEG recording and analysis is in epilepsy, for both clinical monitoring and research studies. For an epileptic brain, the EEG is extremely informative about the disorder during a seizure as well as in between two successive seizures. A large number of quantitative techniques for EEG analysis that we use today were either invented or used on EEGs for the first time to study special characteristics of EEGs of epilepsy patients. One of the first uses of electronic computers to study EEGs was for automatic seizure detection.

6.1 Automatic Seizure Detection

It is well recognized that scalp EEG is much more noisy than intracranial or depth EEG (also known as electrocorticogram or ECoG, collected through intracranial subdural electrodes touching on the cortical surface). For human subjects, the signal-to-noise ratio (SNR) in depth EEG is more than 20 times higher than the SNR in scalp EEG (Table 6.1). Also, depth EEG sources are almost precisely known, whereas scalp EEG sources are quite challenging to determine (see Chapter 3). Naturally, automatic seizure detection is more popular in depth EEG than in scalp EEG (broadly speaking, depth EEG also includes the local field potential [LFP] collected through electrodes penetrating into the cortical layers). Nevertheless, it has been shown that automatic seizure detection in scalp EEG can be quite helpful in monitoring and treating epilepsy (Holmes et al., 2010). The biggest advantages of scalp EEG are that it is noninvasive and it is the easiest and cheapest to acquire among all brain signals. Primarily because of these two advantages, automatic detection of seizures in scalp EEG started soon after the use of electronic computers began to process EEG. By seizure detection we will always mean seizure onset detection (Figure 6.1).

One fundamental difference between communication signal processing and neural signal processing is that the latter has traditionally been processed mostly by visual inspection, whereas communication signals, right from the beginning, are being processed by mathematical, statistical, and

TABLE 6.1

Comparison of SNR between Scalp EEG and Depth EEG (ECoG)

		P1	P2	P3	P4
Prefrontal	SNR EEG	0.08	0.04	0.08	0.11
	SNR ECoG	2.35	4.37	2.51	2.22
	Ratio	29	115	34	21
Motor	SNR EEG	0.57	0.15	0.15	0.52
	SNR ECoG	13.47	26.28	21.19	15.32
	Ratio	24	173	141	30

Source: Taken from Ball, T. et al., *NeuroImage*, 46(3), 708, 2009.
Notes: P1 to P4 are four patients. Ratio denotes the ratio of SNR between ECoG and EEG.

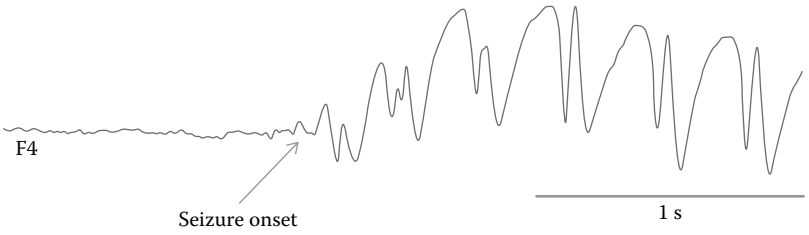


FIGURE 6.1
3-second recording of scalp EEG of an epileptic patient at channel F4. It starts with a normal background EEG followed by a seizure. The seizure onset has been clearly demarcated by a trained epileptologist.

computational techniques (in short, quantitative techniques). Although many of these techniques have found widespread use in neural signal processing as well, processing by visual inspection is still regarded as the gold standard in many aspects of neural signal processing. One notable example of this is the detection of seizure onset and offset. Any automatic detection of seizure onset by quantitative techniques must conform to the onset detected by certified epileptologists who in turn determine seizure onset by visually identifying spikes and sharp waves associated with seizures. Visual identification of seizure onset by certified epileptologists is the gold standard of seizure detection (the average of all the onset points identified by multiple epileptologists of a single seizure is taken as the onset of that seizure). However, the following major challenges are involved (Qu and Gotman, 1997):

1. Seizure onset patterns are highly variable from one patient to another.
2. Seizure patterns may take a long time (more than 20 s) to build up from the background EEG, leading to almost uselessly late detection.
3. The seizure onset pattern of one patient may be very similar to a non-seizure pattern of another patient, especially when artifacts are involved.

This is why seizure onset detection is more difficult than the prominent seizure patterns occurring later during the seizure.

The following observations may help design a system for automatic seizure onset detection: In most patients, one or sometimes two types of seizures tend to occur repeatedly. In these cases, the seizures are quite similar to each other, including the seizure onset. For a particular patient, the seizure onset is most often distinguishable from the background EEG, although it may sometimes resemble the background EEG patterns of other patients.

Here we propose to discuss automatic seizure detection under three different paradigms namely, template based, transformation based, and operator based.

6.1.1 Template-Based Seizure Detection

In this section, we will discuss about the template-based automatic seizure detection algorithm proposed by Qu and Gotman on the continuously monitored EEG of epileptic patients (Qu and Gotman, 1997). It was developed as a seizure warning system for clinical applications.

Ideally, an automated detection algorithm should be able to detect a seizure in depth EEG no later than 5 s after the onset. With the improvement in technology it is likely to improve further. First, an EEG epoch or window of appropriate length will have to be selected. A window length of more than 2.5 s (but not exceeding 5 s) for an EEG sample frequency of 200 Hz or more will be appropriate because this will give a window of 500 time points or longer, which can be used for Fourier transform (Qu and Gotman, 1997). Six features of EEG have been identified to be applied on this window in the hope of detecting seizure onsets with as few misses as possible and at the same time preventing as many false detections of seizures as feasible (Ironically, any algorithm with high accuracy for seizure detection is prone to detecting seizures in the EEG recorded during the period when there were no seizures.). In other words, these six features present in appropriate quantities in an EEG epoch will make the epoch (almost) identical to a seizure onset pattern.

6.1.1.1 Feature Extraction

Following are the six steps of (an epileptic seizure's) feature extraction as they are manifested in scalp EEG.

1. *Average wave amplitude in one epoch*: It is defined as

$$AvgAmp = \frac{\sum_{i=1}^{N} Amp_i}{N}, \tag{6.1}$$

where
N is the number of half-waves in one epoch
Amp_i is the amplitude of a half-wave

The amplitude is measured between a trough and the next peak, after small-amplitude waves are eliminated.

2. *Average wave duration in one epoch*: It is defined as

$$AvgDur = \frac{\sum_{i=1}^{N} Dur_i}{N}, \tag{6.2}$$

where
N is the number of half-waves in one epoch
Dur_i is the duration of a half-wave

Computation of this feature involves filtering out of the small amplitude waves (possibly by thresholding).

3. *Coefficients of variation of wave duration in one epoch*: It is defined as

$$COVA = \frac{\sum_{i=1}^{N} \left(Dur_i - AvgDur\right)^2}{N \times AvgDur^2}, \tag{6.3}$$

where
N is the number of half-waves in one epoch
Dur_i is the duration of a half-wave
$AvgDur$ is the average duration of half-waves in one epoch

This feature reflects the regularity of the duration of half-waves in one epoch.

4. *Dominant frequency*: To every peak in a (Fourier) spectrum of the EEG epoch peak frequency we define two other frequencies in relation to this peak: one is within the rising slope and the other is within the falling slope, and they correspond to amplitudes equal to half the amplitude of the peak. These two frequencies define a frequency band called full-width half-maximum band of the peak. Among all peaks in a spectrum, the peak that has the largest average power in its full-width half-maximum band is called the dominant peak. The dominant frequency is defined as the peak frequency of the dominant peak. This feature is significantly different from the second feature (average duration), as can be illustrated by an example: if the spectrum is dominated by two peaks, one somewhat larger than the other, the mean frequency (as reflected by the average duration) will fall in between these two peaks, whereas the dominant frequency will be the frequency of the largest peak.

5. *Average power in the main energy zone*: The main energy zone is a frequency band that is centered at the average frequency:

$$AvgFreq = \frac{\sum_{i=1}^{N} POWER_i \times freq_i}{\sum_{i=1}^{N} POWER_i} \tag{6.4}$$

   contains 80% of the total energy of a spectrum. The average power in the main energy zone is obtained by dividing the power in this band by its width. It is used to reflect the concentration of energy in a spectrum. If the power in a spectrum is concentrated in one area, the main energy zone is narrow and the average power within it is large. This feature is significantly different from the second feature because this feature is about average power and the second feature is about average duration.
6. *Location feature*: This feature contains the positions of electrodes where a seizure onset occurs. It is basically a list of EEG channels involved in a seizure onset template and therefore there is no unit associated with it. This list can include any number of channels. The list is to be selected by visual observation of the seizure onset by a certified epileptologist. The method attempts to detect seizures with similar patterns in the same brain regions. This characteristic is very important in the definition of a seizure (which is defined as a hypersynchronous activity (Fisher et al., 2005), and therefore adjacent electrodes placed over seizure onset zone are likely to have similar EEG patterns to some extent). In the classification procedure, the location feature is translated into a requirement that the seizure onset occurs in the same channels as those of the template seizure. This feature imposes spatial restraints in the detection method.

6.1.1.2 Representation of Seizure Onset Patterns

A template is represented as a set of points in the detection space. The template is divided into epochs of duration mentioned in Section 6.1.1.1 above, and each point represents one epoch of EEG in one channel. Since the method compares EEG patterns epoch by epoch, the division into epochs of the template, and that of an EEG recording from a channel to be searched for occurrence of a seizure, in which we are looking for a match to the template, affects the probability of finding a match. For instance, as shown in Figure 6.2, pattern A is a template seizure and is divided into epochs from the beginning of the seizure. Pattern B is a new seizure that is randomly divided into epochs, since the beginning of the seizure is unknown at the time of the search. In this case, the epochs in the two seizures are not well aligned, and therefore, a match may not be

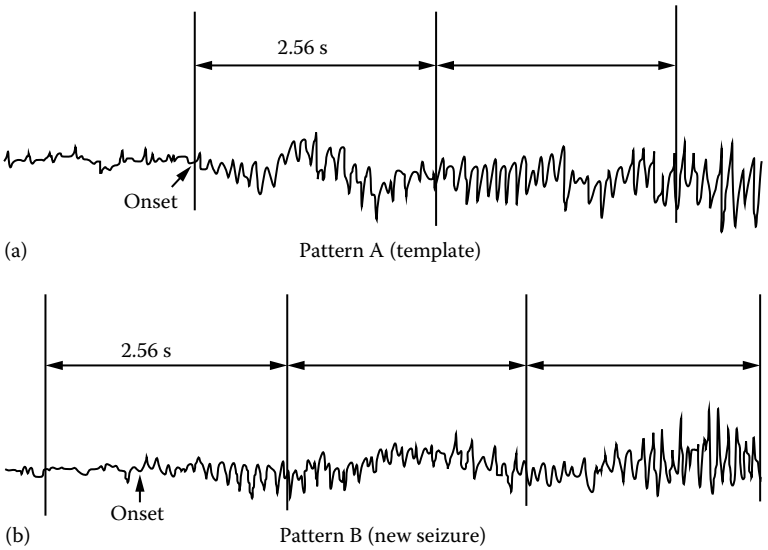


FIGURE 6.2
Epoch division and time alignment: (a) pattern A is a template and (b) pattern B is an unknown EEG segment possibly containing a seizure, which has to be detected by matching with the template of pattern A. The template and the unknown EEG are divided into adjacent epochs resulting in a time alignment problem and a likely mismatch. (Taken from Qu, H. and Gotman, J., *IEEE Trans. Biomed. Eng.*, 44(2), 115, 1997.)

found or may be found late after the onset, despite the similarity of these two seizures. If pattern B was divided into epochs starting at the beginning of the seizure, a match could be made correctly. Time alignment between the template seizure and a new seizure about to be detected is therefore important.

The time alignment problem could be solved by dividing the template or a new seizure into overlapping epochs starting at every sample (i.e., sliding the epoch or the window one point at a time). At least one pair of epochs would then necessarily fall into the best possible alignment. This is quite feasible with a modern desktop workstation in case the signal sample frequency is not too high (say 1 kHz or less). For slower computers the division of the template into epochs can be done by a compromise between the computational load for online processing and the precision of time alignment. In the example shown in Figure 6.3, the template seizure has been divided into epochs starting every 320 ms for an EEG sample frequency of 200 Hz and epoch length 2.56 s or 512 time points. 320 ms is a convenient value because it corresponds to 64 time points, which was the basic processing unit in the program. There were two reasons for the choice of 320 ms. (1) For the template seizure and a new seizure, a mismatch is unlikely to occur when two epochs have a time alignment difference of 320 ms, because EEG changes occurring within 320 ms rarely alter dramatically the

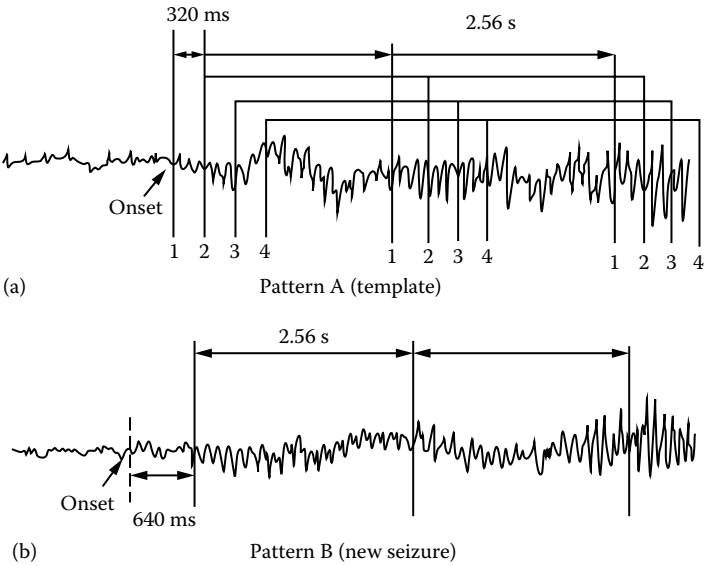


FIGURE 6.3
(a) Division of the template into epochs starting every 64 samples, or 320 ms. In this case, even if at the division of epochs in (b) the new seizure does not start at the onset, a good match still occurs between the first epoch of the new seizure and the third division of the first epoch in the template because both epochs start 640 ms after onset. (Taken from Qu, H. and Gotman, J., *IEEE Trans. Biomed. Eng.*, 44(2), 115, 1997.)

statistical characteristics of a 2.56 s epoch. (2) In comparison with the division of epochs starting every sample, this reduces the number of template points by a factor of 64, thus reducing by the same factor the processing time for classification. This helps in accommodating the limited computational power of older or smaller computers.

6.1.1.3 Distance Measure

Let us note that detection space is a 5-dimensional feature space (because there are five features, (6.1) through (6.4), which are different metrics; the sixth or the channel location feature, which is not a metric, will be used separately), where each epoch is uniquely represented by a 5-dimensional vector. The detection space contains a seizure template as a collection or a cluster of 5-dimensional vectors. There are precisely as many vectors in the seizure template cluster as there are epochs in the whole template. For efficient seizure onset detection the seizure template should be an EEG pattern not extending too far beyond the seizure onset time. We need a distance measure between two points in the detection space, so that the distance of a freshly surveyed epoch of an EEG signal currently being scanned for the detection of possible seizure onset can be measured from the seizure

template points. Let **A** and **B** be two points in the detection space. The distance D between them is defined as

$$D(\mathbf{A},\mathbf{B}) = \left\| (\mathbf{A}-\mathbf{B})\mathbf{M}^{-1}(\mathbf{A}-\mathbf{B})^{T} \right\|, \tag{6.5}$$

where **M** is a 5×5 weight matrix, whose diagonal elements are the weights assigned to the features (6.1) through (6.4). All other entries of **M** are zero. The weight of a feature will depend on the importance of that feature in an epoch in order to make that epoch a seizure template. Once the weights are fixed for a particular distance measure D, they remain unchanged for all detection space calculations. Qu and Gotman arbitrarily took the weights to lie between 1 and 4 (Qu and Gotman, 1997).

6.1.1.4 Onset Detection

In Figure 6.4, it is clear that the distance of a nonseizure point will have to be calculated from the seizure template, which is a set of points (collection of all circles). For an arbitrary nonseizure point **N**, it is done by measuring the distance D between **N** and each of the points of the seizure template and taking the shortest such distance. In Figure 6.4, d denotes the shortest

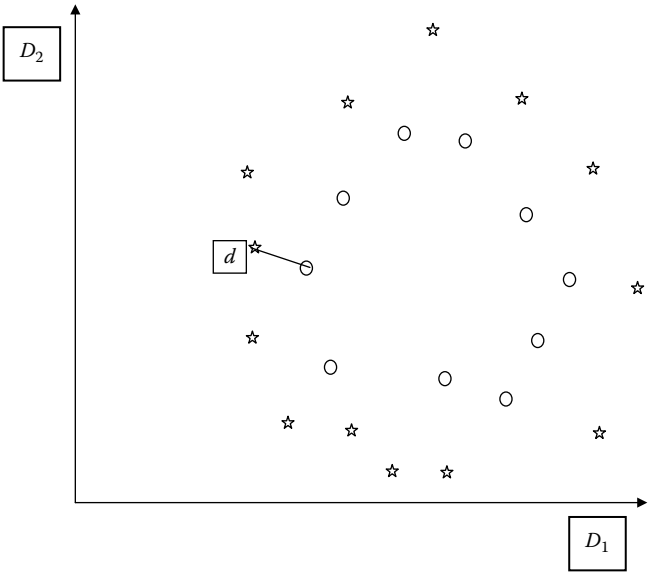


FIGURE 6.4
A 2-dimensional projection of detection space, in which D_1 and D_2 are two independent features. A circle denotes a point belonging to seizure template and a star denotes a point belonging to nonseizure. The shortest distance d between seizure template (cluster of circles) and a nonseizure point has been indicated by drawing a line of shortest length between two such points.

distance across all the nonseizure points from the seizure template. d is obviously a unique quantity, which is fixed by computation taking a substantial number of nonseizure sample points alongside the seizure template. Determining this d is the most crucial part of the classification between seizure and non-seizure in an unknown EEG signal recorded from an epilepsy patient. d is patient specific, because the seizure template is from the particular patient whose seizure onset detection in his/her arbitrary EEG signal is to be accomplished. Nonseizure samples should also be collected from that patient alone, because it is possible that one patient's seizure EEG looks like another patient's nonseizure EEG. We summarize the seizure onset detection rule as follows:

$$\begin{aligned} &\mathbf{U} \text{ is a possible seizure onset if } \left\{\min D(\mathbf{U},\mathbf{X}),\mathbf{X} \in S\right\} < T \times d \\ &\mathbf{U} \text{ is NOT a seizure onset if } \left\{\min D(\mathbf{U},\mathbf{X}),\mathbf{X} \in S\right\} \geq T \times d, \end{aligned} \tag{6.6}$$

where $\mathbf{U}$ is an arbitrary EEG epoch being tested to ascertain if it contains a seizure onset point, S is the seizure template in which $\mathbf{X}$ is an arbitrary point, and T is a threshold taken to be 0.5 by Qu and Gotman (1997).

For utilizing the sixth dimension or the sixth feature, that is, the channel location, the above calculation in the 5-dimensional detection space is carried out at each channel in the focal region (i.e., where the seizure is originating from, assuming it to be a focal seizure; for generalized seizure all the scalp channels are to be considered), and if in 50% of the channels seizure is detected, only then it is taken to be a true detection; otherwise it has to be counted as a false positive.

6.1.2 Transformation-Based Detection

It is evident from Figure 6.1 that time–frequency analysis could perhaps be an effective way for seizure onset detection. Time–frequency analysis can be described as a study of how transiently amplitude and frequency of a signal vary. For a periodic signal time–frequency analysis can be carried out by Fourier transform (see Appendix A). But brain signals hardly repeat themselves at a fixed time lag (In other words, brain signals are nonstationary.). So, Fourier transform, although giving many interesting results on EEG signals, which are also experimentally well established, is not an appropriate tool for EEG analysis from a theoretical point of view. Wavelet transforms (see Appendix B) are more suitable for time–frequency analyses of EEG signals. In this section, we will discuss about how suitable wavelet transforms can result in seizure detection with reasonable accuracy.

The two most important wavelets to be applied on EEG are Morlet and Daubechies wavelets. In Chapter 5, we have seen how to extract the phase of an EEG signal by applying Morlet wavelet transform on it. Similarly, Daubechies wavelet transforms have been found to be suitable for extracting characteristics

of epileptic seizures from EEG signals (Adeli et al., 2003). Due to its clarity and detailed exposition, here we will describe the method following Khan and Gotman (2003), although it was originally meant for depth EEG signals.

We will learn wavelet feature extraction from a single channel seizure EEG signal by an example taken from Khan and Gotman (2003). In the top panel of Figure 6.5 a segment of a single channel (depth) EEG signal s during seizure has been shown. Here, the sample frequency is 200 Hz. So, according to Nyquist criterion, only up to a frequency of $200/2=100$ Hz can be studied in this signal. Now we will apply multiresolution analysis on s in a way that implements Daubechies 4 or D4 (also known as DB4 in the literature) wavelet transformation on s (for more details, see Appendix B).

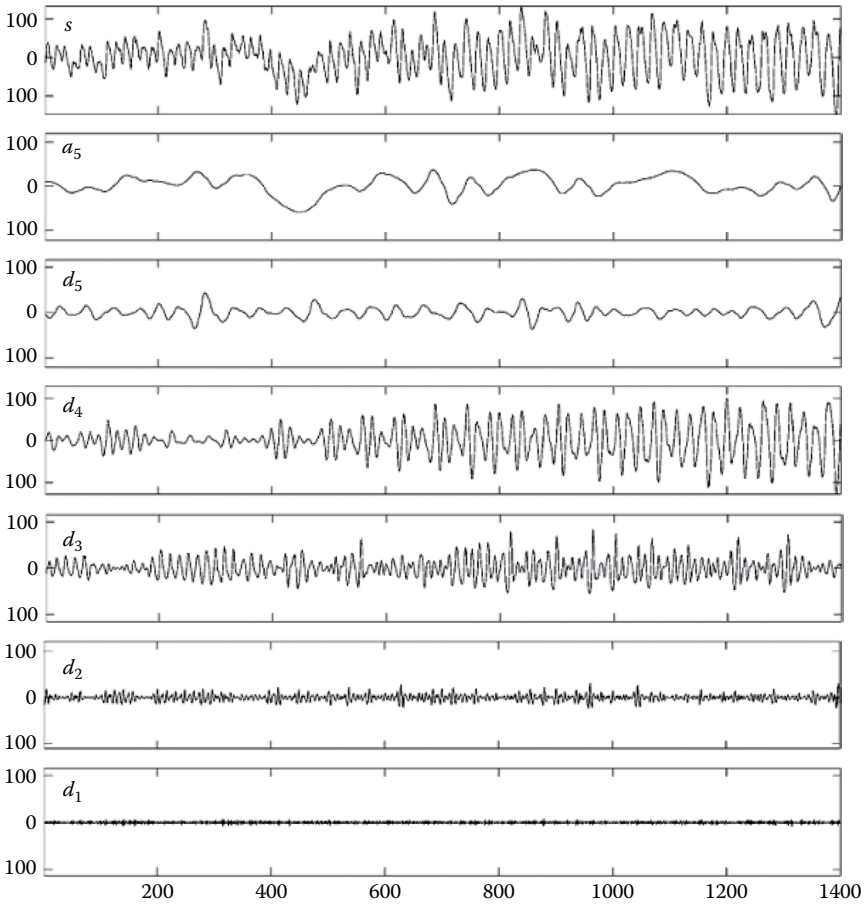


FIGURE 6.5
s is the original EEG signal. d_i is the detailed signal at the ith scale. a_j is the approximate signal at the jth scale. In this particular case, $s=d_1+d_2+d_3+d_4+d_5+a_5$. The abscissa is in time points and the ordinate is in amplitude (arbitrary unit). Here the sample frequency is 200 Hz. (Taken from Khan, Y.U. and Gotman, J., *Clin. Neurophysiol.*, 114, 898, 2003.)

Now the signal will be low-pass filtered and high-pass filtered at a 50 Hz cutoff frequency. The high-pass filtered signal is denoted by d_1, the detailed signal at the first scale. The low-pass filtered signal is denoted by a_1, the approximate signal at scale 1. The approximate signal a_1 was further subjected to high-pass and low-pass filtering at 25 Hz cutoff frequency. The high-pass filtered signal is denoted by d_2, the detailed signal at scale 2. The low-pass filtered signal is denoted by a_2, the approximate signal at scale 2. The approximate signal a_2 was further subjected to high-pass and low-pass filtering at 12 Hz cut off frequency. The high-pass filtered signal is denoted by d_3, the detailed signal at scale 3. The low-pass filtered signal is denoted by a_3, the approximate signal at scale 3. The approximate signal a_3 is further low-pass and high-pass filtered at cutoff frequency 6 Hz, and so on. This way, with recursive filtering, we obtain d_1, d_2, d_3, d_4, d_5, and a_5 from s (Figure 6.5). We get back s simply by $s = d_1 + d_2 + d_3 + d_4 + d_5 + a_5$. The equality holds because $d_1 + d_2 + d_3 + d_4 + d_5 + a_5 = d_1 + d_2 + d_3 + d_4 + a_4 = d_1 + d_2 + d_3 + a_3 = \cdots = s$. In Figure 6.5, seizure characteristics are the most prominent at scales 3 and 4.

Intuitively, the multiresolution analysis of a signal s can be explained by Figure 6.6. $s = a_1 + d_1$, where a_1 is the approximate signal in the low-frequency band and d_1 is the detailed signal in the high-frequency band. a_1 and d_1 are pure sinusoidal signals. So by decomposing s no other components can

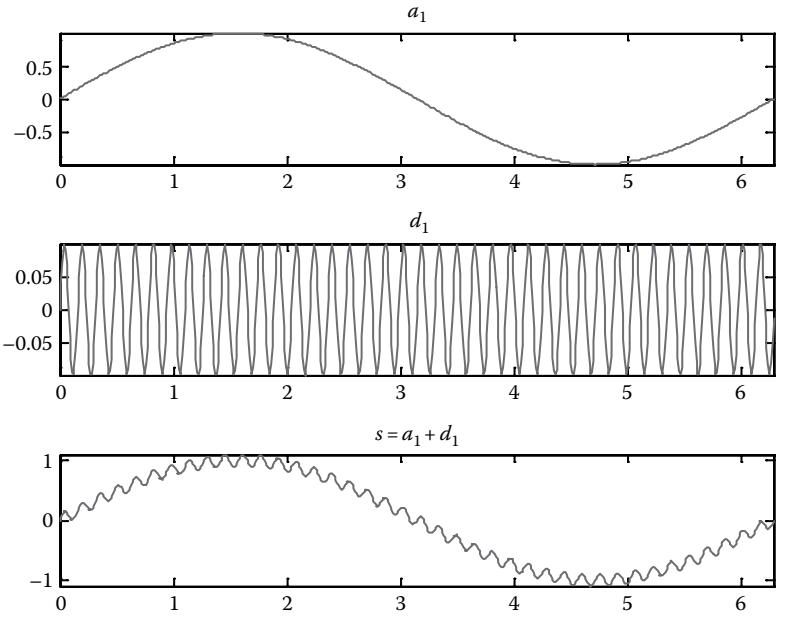


FIGURE 6.6
Multiresolution analysis of signal s (bottom panel) will give the approximate signal a_1 (top panel) and detailed signal d_1 (middle panel) so that $s = a_1 + d_1$ holds. Note that a_1 is a high-amplitude low-frequency signal whereas d_1 is a low-amplitude high-frequency signal.

be obtained. If a_1 has frequency f_1 and d_1 has frequency f_2 ($f_2 > f_1$), then the decomposition of the signal by high-pass and low-pass filtering with cut-off frequency $(f_1 + f_2)/2$ will produce d_1 and a_1, respectively. It is clear from Figure 6.6 that the higher frequency component d_1 takes care of the detailed local characteristics of the signal s, whereas the lower frequency component a_1 decides the global shape or the approximation of the signal s. The signal s is a simple simulated signal, for which one resolution scale is sufficient to decompose it fully into its fundamental components. For a more realistic (and therefore complicated) signal like the one in Figure 6.5, multiple scales of resolution will be necessary. Limit on the number of resolutions is imposed by the sample frequency of the digital signal.

After wavelet decomposition of the single channel EEG signal as shown above, parameters are to be identified on the wavelet transformed components (a_i and d_j) so that the underlying EEG waveform will be recognized as seizure or non-seizure according to the appropriate threshold values of those parameters, which are usually patient specific. For the patient-specific detection scheme, the entire detection system is to be 'trained' on the sample EEG signals of the particular patient. Training will yield the threshold values for the parameters. Three parameters have been mentioned in Khan and Gotman (2003). They are relative energy, coefficient of variation, and relative amplitude.

Relative energy: Energy at the scale l is denoted by $e(l)$ and is expressed as

$$e(l) = \sum_{i=1}^{n} D_i^{2*} \Delta t / n \tag{6.7}$$

where n is the number of discrete wavelet transform (DWT) coefficients (D_i, value of the ith sample of d_l or a_l as the case may be) present in the scale or level l and Δt is the (constant) time interval in between any two consecutive sampled points of the discrete signal. In other words, $\Delta t = 1/Fs$, where Fs is the sample frequency. Relative energy at the scale l is denoted by $e_r(l)$ and given by

$$e_r(l) = \frac{e(l)}{\sum_{i=1}^{n} e(i)}. \tag{6.8}$$

Therefore, for the part of a recording where a seizure occurs, the relative energy of the DWT coefficients for a particular level (scale) rises to a high value and can be detected by setting an appropriate threshold E_{TH} (Khan and Gotman, 2003).

Coefficients of variation: A coefficient of variation is given by σ^2/μ^2, where σ^2 is the variance of the EEG signal (within a window or a segment, see Figure 6.7), and μ is the mean. After wavelet decomposition of the EEG signal, as described above, the coefficient of variation is measured on the wavelet decomposed signal at each level l (see Figure 6.7 for an example) and we denote the coefficient of variation as $c(l)$. A threshold for $c(l)$ for certain

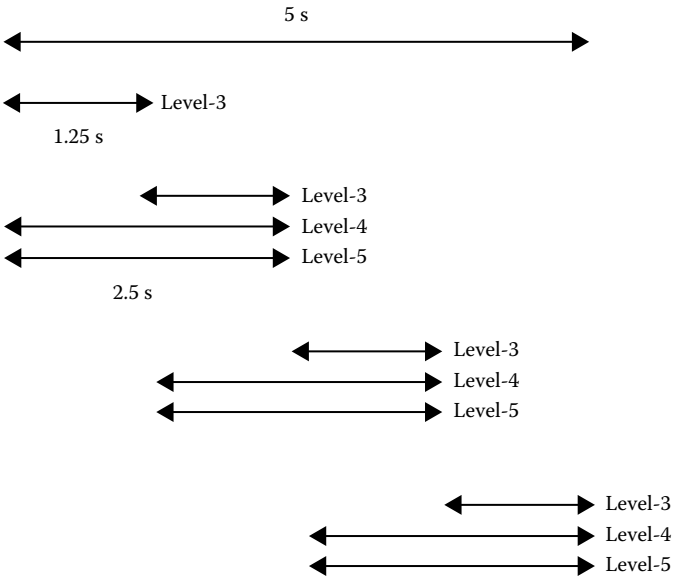


FIGURE 6.7
An example of segmentation of wavelet decomposed EEG signal from a channel. The segments within the 5 s long segment have 50% overlap. (Taken from Khan, Y.U. and Gotman, J., *Clin. Neurophysiol.*, 114, 898, 2003.)

level(s) l needs to be fixed in order to make a distinction between a seizure signal and a nonseizure signal. This threshold is denoted by $C_{TH}(L)$, where L is the level of interest. For an appropriate choice of L, the value of coefficient of variation will be greater than $C_{TH}(L)$ for a seizure signal and will be less for a nonseizure signal.

Relative amplitude: At some point in the course of a seizure, the amplitude of the rhythmic seizure component usually becomes greater than it was in the preseizure background. The average amplitude is computed, as the mean of the amplitudes of the segments. To measure the amplitude relative to the background, each EEG epoch is normalized by the amplitude in the corresponding level in the background. For example, a background of 15 s after leaving a gap of 20 s was selected by Khan and Gotman (2003). The gap is chosen in order to allow detection of seizures having a very gradual onset. The epoch size is varied with the selected level (see Khan and Gotman, 2003 for the implementation). The relative amplitude is then compared with the preset threshold A_{TH}.

6.1.3 Operator-Based Detection

Time–frequency analyses are concerned about how frequency and amplitude of a signal are changing over time. Frequency is modulated by the up–down fluctuation of the amplitude crosspassing a predetermined

horizontal baseline. Derivatives are ideally designed to determine the rate of change of one variable with respect to another. If we measure the rate of change of the signal amplitude with respect to time, we will be able to model many aspects of the time–frequency analyses. For example, epileptic seizures are represented in EEG as spikes and sharp waves. Both spikes and transient sharp changes in wave patterns are accentuated by first- and higher-order temporal derivatives (differentiation with respect to time) of the (time domain) signal. On the contrary, small-amplitude slower waveforms making up the background EEG signals are typically suppressed under the differentiation. In other words, differential operators enhance the contrast of seizure EEG with respect to the background signals thereby facilitating the automatic seizure detection. However, there are two potential bottlenecks to this approach. (1) Temporal derivatives work as a high-pass filter, and thereby it may pick up high-frequency noise, like muscle artifacts, and suppress low-frequency components belonging to the seizure EEG. (2) Since scalp EEG is noisy, the temporal derivative may end up amplifying the noise.

Despite the above two shortcomings various combinations of temporal derivative operators are remarkably efficient in detecting seizure onset in intracranial EEG (Majumdar and Vardhan, 2011; Majumdar, 2012).

Figure 6.8 represents a schematic block diagram of the automatic seizure detection by differential windowed variance (DWV) algorithm. t stands for time, $W()$ stands for window, and $W_{-1}()$ denotes the window immediately preceding the window of maximum variance. The DWV algorithm as presented in Figure 6.8 has three major subparts, namely enhancement (accentuate the seizure part of the signal compared to the background), filtering (DWV also works as a high-pass filter), and detection (which actually identifies the seizure onset point).

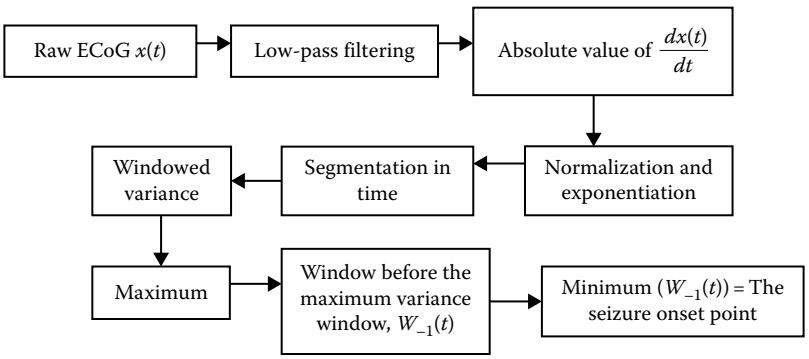


FIGURE 6.8
A schematic block diagram of seizure detection by differential operator and windowed variance (DVW). (Taken from Majumdar, K. and Vardhan, P., *IEEE Trans. Neural Syst. Rehabil. Eng.*, 19(4), 356, 2011.)

DWV achieves the enhancement of seizure features in a single channel brain electrophysiological signal $x(t)$ (ECoG and EEG) in the following way:

1. $y(t) = G(x(t), f)$, where G is low-pass filter and f is the cutoff frequency. Low-pass filtering is necessary for removing unnecessary high-frequency components (seizure characteristics in the EEG remain within 40 Hz), which may get enhanced due to differentiation (for it is a high-pass filter), leading to masking of seizure characteristics.
2. $X(t) = \exp\left(\frac{1}{w}\left|D'y(t)\right|\right)$, D' is the derivative with respect to t and w is a normalization constant (positive and large, it was taken to be $w = 100,000$ in Majumdar and Vardhan, 2011).

Due to the transformation $x \to X$, as described in 1 and 2, in case x is bearing an electrographic seizure, the seizure part of x gets enhanced considerably with respect to the background (which is suppressed to a good extent) in X. The reason is very simple. Let a, b, and c be successive time points. If a spike (peak) occurs at b then statistically $x(b) - x(a)$ and $x(c) - x(b)$ both have high numerical value, which are two successive points in $D'x$. On the contrary, if all the three points belong to normal background signal, then $x(b) - x(a)$ and $x(c) - x(b)$ will both have small values (actually smaller than the average background signal amplitude). Thus $D'x$ will enhance the spikes of x, but will suppress its background. For a very high sampling rate this advantage will be lost, but in that case the difference operation can be carried out as $x(t+k) - x(t)$ instead of $x(t+1) - x(t)$ for a suitably chosen integer $k > 1$.

Windowed variance is a very efficient method to detect changes in an EEG or ECoG signal due to epileptiform activities. At the same time it is a linear method. First, $V_W(X(t), l, s)$ will have to be calculated, where V_W is the windowed variance operation with window length l and sliding distance s. The windowed variance graph $\mathrm{Wvar}(t) = V_W(X(t), l, s)$ has been shown in Figure 6.9 with peak onset and offset clearly identified as the local minimum before and after the seizure peak, respectively. The seizure peak has been automatically identified as the maximum value of $\mathrm{Wvar}(t)$ in a one-hour-long snapshot of continuous stream ECoG data. Once a peak is identified as a seizure, it is marked and never taken into consideration for further identification. The process continues to search for the seizure peak after the last identified one. Please note that although the seizure peak identification is in real time, the actual seizure onset detection is slightly behind the real time.

6.1.3.1 False-Detection Avoidance

The more efficient a seizure detection method, the higher is the chance of false seizure detection in the nonseizure data. In order to eliminate false detection as far as possible, eight empirical tests have been introduced by Majumdar and Vardhan (2011). To be qualified as a seizure the data must be

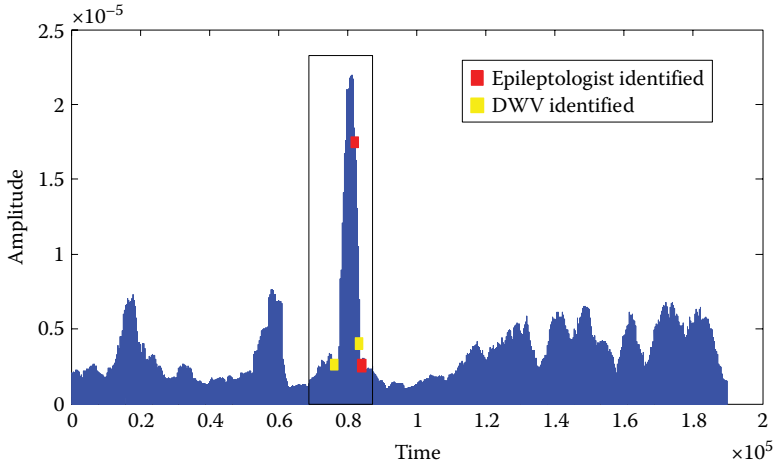


FIGURE 6.9
A Wvar(t) snapshot of an ECoG signal during the seizure from a focal channel, in which the enclosed in the rectangular box is the seizure neighborhood and is distinctly visible due to the DWV steps 1 and 2 above. The epileptologist-identified seizure onset and offset points are 82,010 and 83,855, respectively, and those by DWV are 76,802 and 83,702, respectively. Seizure onset and offset time points are identified by the vertical arms of the small solid rectangles (right arm for the left-most yellow rectangle and left arm for all other rectangles), red for epileptologist-identified points and yellow for DWV-identified points. The abscissa denotes the time in seconds (256 points = 1 s) and the ordinate denotes the normalized amplitude. (Taken from Majumdar, K. and Vardhan, P., *IEEE Trans. Neural Syst. Rehabil. Eng.*, 19(4), 356, 2011.)

identified as having seizure by each of the eight tests separately. Otherwise it has to be classified as a nonseizure.

1. Maximum windowed variance on differentiated signals: The windowed variance operation is to be performed on $X(t)$ for a fixed window length (across all patients) and fixed sliding distance (across all patients). Let the output of this operation be denoted as Wvar($X(t)$) or more simply Wvar(t). The maximum windowed variance across a segment of predetermined length (e.g., 1 hour in Majumdar and Vardhan, 2011) be denoted by B = max(Wvar(t)). For a given patient, if B is below a certain threshold (to be determined by sample seizure data of the patient), then it is a nonseizure. Otherwise it is a seizure. In fact, the value should lie not only above a threshold, but also below another threshold. If it goes above the latter threshold, it is most likely to be an artifact rather than a seizure. So, ideally B should lie in an interval (patient specific).
2. Maximum windowed variance on undifferentiated signal: Let be $y(t)$ as in step 1 of DWV. Maximum windowed variance operation is to be performed on the absolute value of $y(t)$ with the same window length and

sliding distance as in (1) across all patients. The variance of the window of maximum variance has been denoted as C. If C falls below certain threshold (to be determined by sample seizure data of the patient) then it is a nonseizure. Otherwise it is a seizure. Just for the same reasons as in (1) in the case of a seizure C should lie within a patient-specific interval.

3. The window after the maximum variance window: Let Wvar(t) be as in (1). Let the window of maximum variance B be denoted by W_0. Let the window immediately after W_0 be denoted as W_1. Let mm be an array of variance values of successive windows. Let us write $mm(1) = \text{var}(W_1)$, where var( ) denotes the variance. Let us consider $|B - mm(1)|$. If $|B - mm(1)|$ falls below a certain threshold (to be determined by sample seizure data of the patient), then it is a nonseizure. Otherwise it is a seizure. Here too for seizure the value should lie within a patient-specific interval.
4. The 3/4th rule: Let $mm(i)$ be the variance of $X(t)$ at the ith window W_i after W_0. The number of windows to be considered after W_0 is to be stipulated beforehand (it was kept at 16 in Majumdar and Vardhan, 2011). Let $n1(1)$ be the position (taken to be the first or starting point of the window) of the first of the windows with the minimum variance among the stipulated number of windows. Let M is the position of W_0 and $N1 = M + (\text{window length}) * n(1)$. The local minimum of $X(t)$ within W_{-1} has been treated as the *onset* of seizure (Figure 6.9) and the local minimum of $X(t)$ within the window W_1 is taken to be the *offset*. While the onset identification has been satisfactory (matches closely with the judgment of epileptologists) the offset has not been that reliable (Majumdar and Vardhan, 2011). However, it has an important role in distinguishing between seizure and nonseizure peaks, as will be seen in (8) below. Let E be an array consisting of values of Wvar(t) starting from two windows before M, that is, W_{-2} through $W_{n1(1)}$. Let z be an array consisting of the maximum values of E. Let F be another array consisting of values of E that are greater than or equal to $\frac{3}{4}\max(E)$. $||\lfloor mean(F) - z(1) \rfloor||$ is a quantity whose threshold distinguishes between seizure and nonseizure EEG. $\lfloor x \rfloor$ is the least integer greater than x, called *floor*. If it lies below a certain threshold, then it is a seizure, otherwise it is a nonseizure. It is worth mentioning that compared to the other thresholds described in (1), (2), and (3), this threshold took fewer different values across the patient population. In other words, this threshold has remained more uniform across the patient population (Majumdar and Vardhan, 2011).
5. The above 3/4th rule: $\lfloor std(F) \rfloor$, where *std* stands for standard deviation, is a quantity whose patient specific threshold will distinguish between seizure and nonseizure signal windows. There is no seizure if the value is greater than the threshold. This threshold also

remained more uniform across the patient population that we have tested compared to the thresholds in (1), (2), and (3).

6. The double derivative of E (DDE) rule: Let $DE = \exp\left(\frac{1}{v}|D''E|\right)$, where D'' denotes double differentiation and v is a positive normalization constant (kept fixed across all the patients). E is as given in (4). Let K be an array consisting of values of DE, which are greater than or equal to $0.999 * \max(DE)$ (Majumdar and Vardhan, 2011). The length of K is a quantity whose threshold distinguishes between seizure and nonseizure EEG. A K value greater than the threshold length indicates nonseizure. This threshold is also more uniform across the patient population.
7. Maximum DE (MDE) rule: For seizure EEG, $\max(DE)$ must lie within a specified interval. This interval is patient specific.
8. Peak rejection: The difference between the seizure onset and offset points as described in (4) gives the width of a seizure pillar (the tall column enclosed within the boxed part in Figure 6.9). Seizure pillar roughly consists of the entire seizure signal in $X(t)$. In the case of some patients' data where there are too many noisy spikes, this method will help eliminate nonseizure spikes from the signal, thereby making the detection more accurate (Figure 6.10). The noise spikes are usually less wide than the seizure spikes (Figure 6.10). So if the peak width is less than the threshold, it is taken to be a nonseizure peak. Otherwise it is a seizure peak.

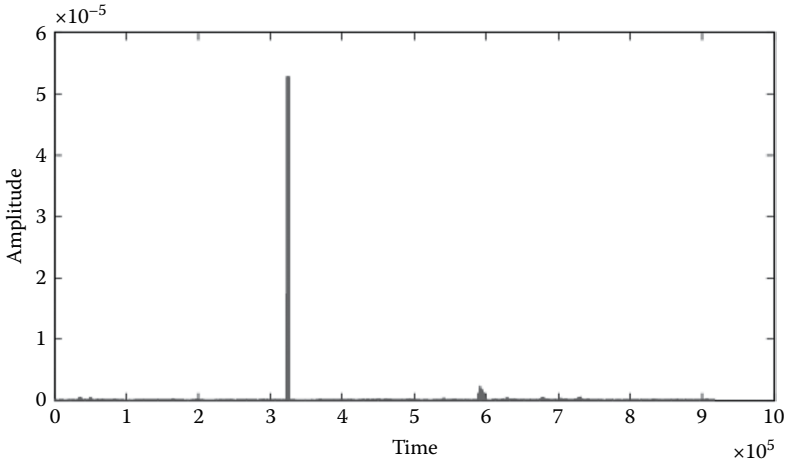


FIGURE 6.10
$X(t)$ form of a seizure ECoG $x(t)$. The seizure peak appears distinctively as a short, thick, dense pillar above the time point 6×10^5. A tall, thin, nonseizure peak appears between 3×10^5 and 4×10^5, which is to be eliminated by the peak rejection method. 256 time points = 1 s. (Taken from Majumdar, K. and Vardhan, P., *IEEE Trans. Neural Syst. Rehabil. Eng.*, 19(4), 356, 2011.)

Note that in the DWV algorithm described above the expression $\frac{1}{w}\sum_{i=1}^{w}\left|D'y_i(t)\right|$ is known as line length (Guo et al., 2010), where w was taken to be the number of sample points in the window. Line length is said to be an excellent feature for high-accuracy seizure onset detection with low false positives and shorter detection delay both in intracranial (Guo et al., 2010) and scalp (Logesparan et al., 2012) EEG. In DWV a more refined form of the line-length feature was used. Notice that both line length and DWV are linear time implementable and are therefore suitable for online implementation. Exponentiation amplifies the contrast between the seizure and background EEG, making the onset detection easier. However, this largely depends on the quality of data. For example, if the EEG signal is contaminated with high-frequency noise during a seizure, both DWV and line-length operation will enhance the noise with respect to the background EEG, leading to false detection. In that case, false detection–avoidance heuristics described above will be useful.

More recently, the absolute value of $X''(t)X'(t)$ has been found to be useful for seizure detection, when the difference has been calculated as the average of the forward and backward difference, that is,

$$X'(n)=\frac{1}{2}\left\{\left(X(n)-X(n-1)\right)+\left(X(n+1)-X(n)\right)\right\}. \tag{6.9}$$

A justification for (6.9) can readily be given from conventional calculus. Let $X(t)$ be a continuous real-valued function of a real variable, which is differentiable at $t=t_0$. Then

$$X'(t_0)=X'(t_0-)=X'(t_0+)=\frac{1}{2}\left\{X'(t_0-)+X'(t_0+)\right\}, \tag{6.10}$$

where $X'(t_0-)$ and $X'(t_0+)$ are the backward and forward derivatives, respectively.

It has been compared with line length and found to be performing marginally better on a publicly available benchmark data set, namely the Freiburg Seizure Prediction Project, 2008 data. This data was acquired from the Epilepsy Center of the University Hospital of Freiburg, Freiburg, Germany, and contains 87 seizures recorded from 21 focal epileptic patients. The efficacy of $|X''(t)X'(t)|$ was tested on a total of 261 focal channels during preictal, ictal, and postictal recording. This is depth EEG data with sample frequency 256 Hz and contains usual artifacts. The seizure onset and offset points were identified by certified epileptologists at the place of acquisition. Also, the focal and nonfocal channels were identified. Seizure detection algorithms working well on this data should work well on the scalp EEG as well. Out of these 261 focal channels seizure onset was detected better by $|X''(t)X'(t)|$ on 103 channels, by line length on 84 channels, and by the plain

raw data on 74 channels. Each channel contained a one-hour recording with exactly one seizure during that time and also preictal and postictal signals. The conclusion was drawn by *receiver operator characteristic* (ROC) curve analysis (see Section 6.5) and comparing the area under the curve (AUC) for all the three different methods mentioned above. But a modified version of the line-length algorithm can perform even better in detecting seizures. Since line length is basically a first-order difference operation, if it can be written as an average of forward and backward differences it works better than the classical line-length algorithm and even better than $|X''(t)X'(t)|$. Moreover, line length being a first-order difference enhances the high-frequency artifacts much less and is therefore less prone to false-positive detection (Pathak et al., 2017).

6.2 Lateralization

In the case of focal seizures, lateralization is the problem of locating the seizure focus either in the left hemisphere or the right hemisphere of the brain. Qualitatively, it can be accomplished through the study of various clinical seizure manifestations (Marks and Laxer, 1998; Bonelli et al. 2007), which are called *seizure semiology* (Noachtar and Peters, 2009). It can also be accomplished with high accuracy through judicious processing of the multichannel scalp EEG signals, without the aid of any information about clinical seizure manifestations (Cecchin et al., 2010). In that sense, it is a special case of seizure source localization problem from scalp EEG signals recorded during the onset of epileptic seizures.

One of the first quantitative methods for lateralization with ictal EEG used frequency information. It was based on the right–left difference in the averaged relative frequency band power: the best results were obtained with the pairs of electrodes T4–T6/T3–T5 and the frequency band 4–10.5 Hz (Murro et al., 1993). In another work, frequency with maximal global field power (GFP) was computed by means of an FFT and then the epoch showing the highest GFP allowed localizing the epileptic activity with a linear inverse solution algorithm (Blanke et al., 2000). An epileptic abnormality index (EAI) was constructed from frequency band power and power asymmetry (left and right) parameters and used for EEG background activity analysis with 17 electrode recordings performed during wakefulness with eyes closed. This EAI allows lateralizing epileptic abnormalities, but EEG parts containing excessive artifacts were visually rejected before analysis (Temuçin et al., 2005). van Putten and coworkers defined a brain symmetry index (BSI), which was originally designed to assist visual interpretation of EEG during carotid endarterectomy (surgical removal of blood clot from the carotid artery for the prevention of stroke), to detect temporal lobe seizures and consequently became useful for

lateralization (van Putten et al., 2005). This BSI is the absolute value of the relative interhemispheric difference in spectral density, estimated from the mean of eight left–right symmetric bipolar EEG channels. They also defined phase synchronization-related symmetry measures. An improved version of the BSI was proposed later (van Putten, 2007).

The electrical activity of the brain has particular symmetric relations. In particular, the spectral distributions of bipolar EEG recordings from bilateral symmetric positions on the two cerebral hemispheres are nearly similar under normal physiological conditions. In various pathological conditions, however, hemispheric asymmetries can arise, for example, during ischemia. This may be detected by changes in the symmetry properties of the electroencephalogram. This phenomenon is the basis for *brain symmetry index* (BSI, van Putten et al., 2005).

There is a large group of paroxysmal (i.e., cause not known) neurological conditions, like focal seizures, that may induce asymmetries in the electroencephalogram. This property is often detected by visual inspection of the EEG in several patients suffering from, for example, temporal lobe epilepsy (TLE). This is the motivation behind proposing BSI for lateralization of seizure focus from scalp EEG recording. The BSI is defined as

$$BSI(t) = \frac{2}{MN} \sum_{k=1}^{M} \sum_{i=1}^{N/2} Q_{ik}(t), \tag{6.11}$$

where $Q_{ik}(t)$ is given by

$$Q_{ik}(t) = \left\| \frac{R_{ik}(t) - L_{ik}(t)}{R_{ik}(t) + L_{ik}(t)} \right\|, \tag{6.12}$$

where

N is the number of bipolar channels

M is the number of different frequencies considered (van Putten et al., 2005)

Imagine a symmetric placement of EEG electrodes on two hemispheres of the head (Figure 6.11). Odd-numbered signals are coming out from the left hemisphere and even-numbered signals are from the right hemisphere. Figure 6.11 with the caption describes how to calculate Q_i for each channel i, with an example for $i = 1$. Then $BSI(t)$ is calculated. Actually, $BSI(t)$ is calculated over an interval to which t belongs. In this sense, in Figure 6.11 also t in $L_{ik}(t)$ is actually representative of an interval, not a time point. Note that the lower bound for the BSI is zero (perfect symmetry), whereas for the upper bound $BSI = 1$, which implies maximal asymmetry. Under normal conditions $BSI \approx 0.05$ (van Putten et al., 2005).

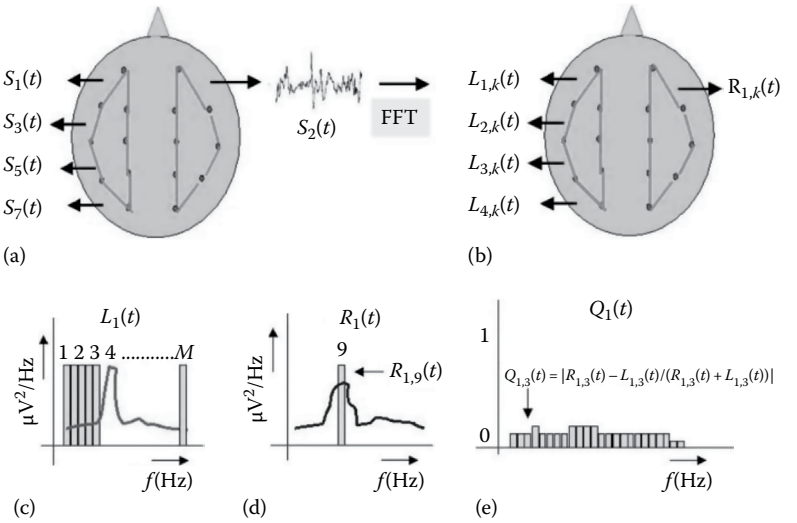


FIGURE 6.11
(a) Symmetric placement of EEG electrodes on both the hemispheres of the head. Odd-numbered signals are from the left hemisphere and even-numbered signals are from the right hemisphere. (b) Fourier-transformed signals—S_1 becomes $L_{1,k}$ after taking the square of absolute value of the kth Fourier coefficient, S_2 becomes $R_{1,k}$ after taking the sum of squares of absolute values of the Fourier coefficients, and so on. (c) Signal power density of S_1 as sum of $L_{1,k}$ across all k divided by the total number of frequencies. (d) Signal power density of S_2 as sum of $R_{1,k}$ across all k divided by the total number of frequencies. (e) Q_1 calculated (across all frequencies) according to Equation 6.12. (Taken from van Putten, M.J.A.M. et al., *Clin. Neurophysiol.*, 116, 2480, 2005.)

6.3 Interictal EEG

In patients with epilepsy not only the EEG during a seizure but also the EEG recorded in between seizures is informative and worth analyzing for diagnosis and prognosis. The EEG recorded in between two successive seizures is called *interictal* EEG. However, scalp EEG recording is subject to certain limitations that may prevent detection of interictal spikes. Scalp electrodes sample only one third of the cortex, so that discharges arising within sulci, in basal regions (e.g., orbitofrontal cortex), and in interhemispheric regions (e.g., interhemispheric supplementary motor cortex) are not detected. Epileptiform activities generated by buried cortex such as amygdala and hippocampus may not be captured on scalp recordings. Bone, dura, and scalp tissue attenuate EEG signals, further hampering the sensitivity of scalp recordings. The orientation of a spike dipole (dipole source of the spike) must be orthogonal to the surface (or nearly so), and dipoles that are parallel to the scalp cannot be detected by scalp electrodes. A large area of cortex, approximately 6 cm^2, must be involved in a spike discharge for that spike to be apparent for scalp recording.

Scalp recordings are subject to influence of muscle and other artifacts, which may obscure the interictal spikes in EEG (Pillai and Sperling, 2006).

Diagnostic uses of interictal EEG includes the following: (1) aiding to establish whether epilepsy is present, (2) determining whether the seizure is focal or generalized (a seizure is *focal* when originated from small, pointlike region in the brain, and it is *generalized* when it originates simultaneously from a large part of the brain) in nature, and (3) assisting in defining specific syndromes, for example, rolandic epilepsy, childhood absence epilepsy, and Lennox–Gastaut syndrome (Pillai and Sperling, 2006). Lennox–Gastaut syndrome is a type of epilepsy with multiple different types of seizures.

Interictal EEG contains spikes and sharp waves, which makes it different from the normal background EEG. These interictal spikes and sharp waves are also known as *interictal epileptiform discharge* (IED) and are a typical feature of EEG recorded from epileptic patients when not having seizure (but in some cases it may also be observed in individuals with no history of epilepsy, see So, 2010). Interictal EEG is characterized by the following criteria (Pillai and Sperling, 2006): (1) it should be paroxysmal (sudden increase in the amplitude of the EEG signal) and clearly distinguished from the background activity; (2) it typically shows an abrupt change in polarity occurring over several milliseconds resulting in a sharp contour or spikiness; (3) its duration should be less than 200 ms, with spikes lasting between 20 and 70 ms, and sharp waves lasting between 70 and 200 ms, the distinction is morphologic in nature, and there is no clinical reason to distinguish between them; (4) it must have a physiologic field, with the discharge recorded from more than one electrode and a voltage gradient should be present; (5) spikes or sharp wave are typically negative in polarity; and (6) most spikes are followed by an after-going slow wave.

Automatic detection of spikes and sharp waves (SSW) in interictal EEG is a challenging problem. Here, we will present an algorithm for it following Gotman and Gloor (1976). First, the signal is divided into segment- and sequence-bearing features of SSW, and then it is further refined by rejecting SSW-like artifacts. Let us first briefly describe below some of the important and quantifiable attributes of SSW. For more details the original paper should be consulted.

1. Dividing an original EEG signal into segments and sequences has been elaborated in Figure 6.12. SSWs are identified by duration, amplitude, and direction in the segmented EEG.
2. Waves have been described in Figure 6.12c. Simple rhythmic activities, such as delta and alpha waves, are made up of sequences consisting of single segments.
3. In order to separate out the SSW from the background, the latter was taken to be the average signal over a 5 s window with 1/3 s sliding. The ratio of maximum amplitude of a wave and that of the preceding 5 s window is the relative amplitude of the wave whose value above a (patient-specific) threshold may be indicative of a SSW

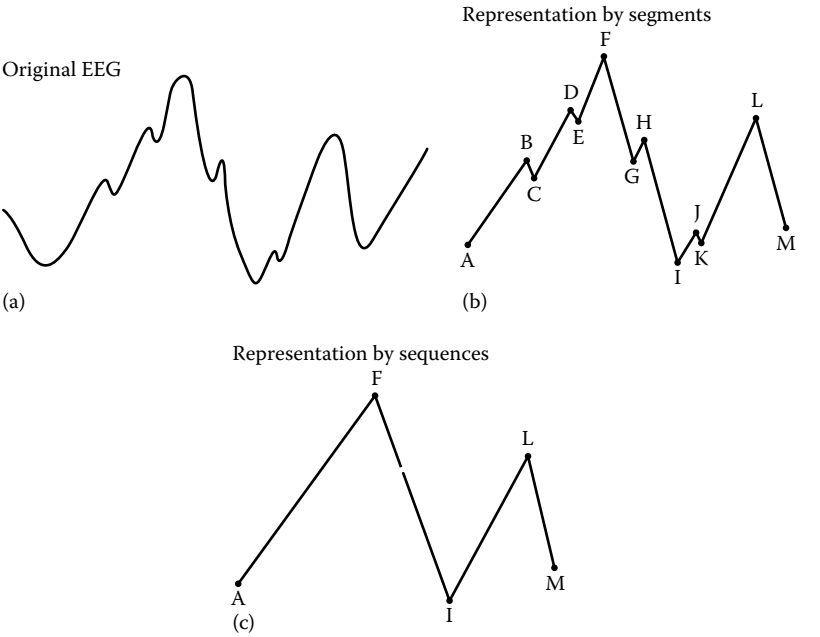


FIGURE 6.12
(a) EEG signal. (b) The signal in (a) has been approximated by straight line *segments*, each of which is joining two consecutive extrema of the signal in (a), for example, AB, BC, CD, DE, GH, HI, etc. (c) A *sequence* is a collection of segments belonging to a monotonically increasing (or decreasing) part of the signal, possibly with a number of isolated exceptions, for example, AF is a sequence of monotonically increasing chains of segments AB, CD, and EF, with BE and DE being decreasing as exceptions. A *wave* is a collection of two adjacent sequences, one is increasing and the other is decreasing, for example, AFI and FIL in (c). (Taken from Gotman, J. and Gloor, P., *Electroencephal. Clin. Neurophysiol*., 41, 513, 1976.)

complex (in Gotman and Gloor, 1976; this value was reported to lie between 0 and 20). In fact, the 5 s long window was divided into 16 sub-windows of 320 ms duration each ($16 \times 0.320 = 5.12$ s). If in a 320 ms sub-window there are N sequences, then the average amplitude of that sub-window was calculated as $A/(N + 6)$, where A is the sum of the amplitude of all the N sequences. Additive 6 in the denominator is a bias term. The average amplitude across the 5 s window is the average across all the 16 values of the form $A/(N + 6)$.

4. Since SSW has a sharpness component, to aid detection it should be enhanced compared to the background. One of the most straightforward ways to enhance it is to perform the temporal difference operation on the digitized EEG. The second difference operation was chosen by Gotman and Gloor (1976) for this purpose. The relative second difference was calculated in exactly the same way with a 5 s window as the relative amplitude in point 3 above.

5. The pseudo-duration of a half-wave (a wave consists of two sequences, each of them is a half-wave) is elaborated in Figure 6.13. Although both the half-waves in the Figure 6.13 have the same amplitude and duration, their pseudo-durations are different. Half-wave ocb has a much larger pseudo-duration and is therefore unlikely to be part of SSW.
6. Two more final features of EEG according to Gotman and Gloor (1976) concern the third half-wave (Figure 6.14). The relative amplitude and duration of the third half-wave are the two features that can distinguish SSW from the background EEG.

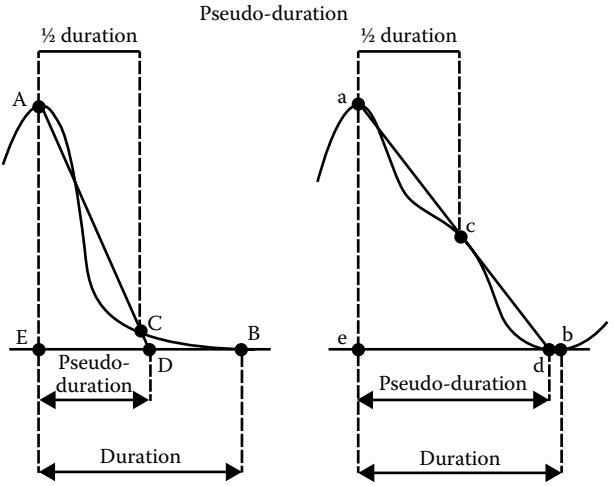


FIGURE 6.13
ACB and ocb are half-waves. C is a point on the half-wave ACB, such that starting from A up to C the wave traverses half of the half-wave duration. EB is the baseline. The line AC cuts EB at D. ED is the pseudo-duration of the half-wave. Similarly, for ocb ed is the pseudo-duration. (Taken from Gotman, J. and Gloor, P., *Electroencephal. Clin. Neurophysiol.*, 41, 513, 1976.)

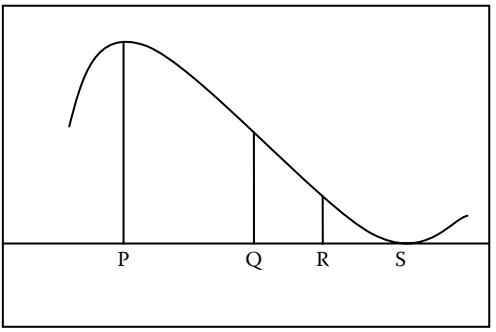


FIGURE 6.14
PS is the duration of the half-wave, QS is the duration of the second half-wave and RS is the duration of the third half-wave.

The channels are processed independently of each other. The signal is decomposed into segments and sequences. When a segment or a sequence is complete, the following operations take place sequentially (if the answer to any question is 'No', analysis proceeds to the next half-wave):

1. Is the relative amplitude of the current half-wave and of the preceding one above the preset threshold? A relative amplitude of 4 is necessary for a half-wave to be considered as a possible part of SSW.
2. Are the pseudo-durations of both half-waves under the appropriate limit? In order to belong to a SSW complex a half-wave must have its pseudo-duration and relative amplitude satisfying the requirements plotted in Figure 6.15. The region ABCD is the region of validity for a first half-wave and region AEFGH is for a second half-wave.
3. Given the relative sharpness of a wave, are the relative amplitudes large enough? A wave of small relative sharpness is a SSW complex only if its relative amplitude is large enough. A wave of larger relative sharpness requires a smaller relative amplitude to be identified as a SSW complex. Table 6.2 indicates the thresholds of relative amplitude of the first, second, and third half-waves for a given relative sharpness. Table 6.2 is to be interpreted as follows: To be a SSW complex, a wave must have a combination of its 3 relative amplitudes in agreement with at least one horizontal line in the section of its relative sharpness.

   In Table 6.2 the threshold value 4 has been subtracted from the relative amplitude, leaving values in excess of the threshold.

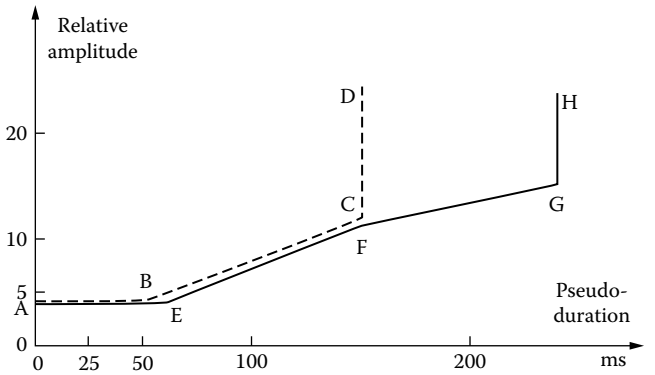


FIGURE 6.15
Diagram of the limits of the relative amplitude as a function of the pseudo-duration. Region ABCD is for a first half-wave and region AEFGH is for a second half-wave. The coordinates are A (0,4), B (50,4), C (150,12), D (150,–), E (60,4), F (150,11), G (240,14), and H (240,–). (Taken from Gotman, J. and Gloor, P., *Electroencephal. Clin. Neurophysiol.*, 41, 513, 1976.)

TABLE 6.2

Relative Sharpness and Relative Amplitude Thresholds for SSW Detection

Sharpness	Amplitude of First Half-Wave	Amplitude of Second Half-Wave	Amplitude of Third Half-Wave
0	≥3	≥4	—
1	1	≥4	—
	≥2	≥3	
2	0	1	≥4
	1	1	≥2
	0	2	≥2
	0	≥3	—
	≥1	≥2	—
≥3	0	0	≥4
	0	1	≥2
	0	2	—
	1	1	—
		Sum ≥ 3	—

Source: Taken from Gotman, J. and Gloor, P., *Electroencephal. Clin. Neurophysiol.*, 41, 513, 1976.
Note: The absolute threshold 4 was subtracted from the relative amplitude.

It may be noticed that (a) waves of high sharpness (≥2) with first and second half-waves of small amplitude require a third half-wave of large-enough amplitude. (b) Thresholds for the second half-wave are often larger than those for the first, reflecting the frequent asymmetry seen in SSW.

4. Is the total duration of the wave is larger or equal to a minimum value (35 ms)?

If answer to all the above questions is 'Yes', the wave under investigation is declared to be a possible SSW complex.

6.4 Seizure Prediction

Epileptic seizures give rise to distinct patterns in the EEG. These seizure patterns in the EEG are called *electrographic seizures*. The clinical onsets of seizure and electrographic seizure may not always coincide, but for most of the time they do. An interesting question is, 'Is there a pattern in EEG as a precursor of an epileptic seizure?' If yes, how long in advance of a seizure onset does that pattern occur? How to identify that pattern? These are some of the questions haunting the EEG community ever since electrographic

seizures have been discovered. They are studied under the discipline 'seizure prediction'. Many quantitative techniques have been tried for automatic seizure prediction in EEG by electronic computers only with moderate success (for a review, see Mormann et al., 2007). Nevertheless, some of those techniques are worth discussing under the broad title of 'quantitative EEG'.

Two-thirds of about 1% epileptics of the world's population achieve sufficient seizure control from anticonvulsive medication, and another 8%–10% could benefit from resective surgery. For the remaining 25% of patients, no sufficient treatment is currently available. For epilepsy patients who do not achieve complete seizure control, it is the sudden, unforeseen way in which seizures strike 'like a bolt from the blue' that represents one of the most disabling aspects of the disease. Apart from the risk of serious injury, there is often an intense feeling of helplessness that has a strong impact on the everyday life of a patient. A method capable of predicting the occurrence of seizures could significantly improve the therapeutic possibilities and thereby the quality of life for epilepsy patients (Mormann et al., 2007).

If it were possible to reliably predict seizure occurrence from dynamical changes in the EEG of epilepsy patients, fully automated closed-loop seizure-prevention systems could be envisioned (Closed-loop seizure suppression by electrical stimulation has already been realized and got FDA (US Food and Drug Administration) approval in 2013). Treatment concepts could move from preventive strategies (e.g., long-term medication with anti-epileptic drugs) toward an EEG-triggered on-demand therapy, for example, by excretion of fast-acting anticonvulsant substances or by electrical or other stimulation in an attempt to reset brain dynamics to a state that will no longer develop into a seizure (Mormann et al., 2007).

In principle, there are two different scenarios of how a seizure could evolve (Lopes da Silva et al., 2003). It could be caused by a sudden and abrupt transition, in which case it would not be preceded by detectable dynamical changes in the EEG. Such a scenario would be conceivable for the initiation of seizures in primary generalized epilepsy. Alternatively, this transition could be a gradual change (or a cascade of changes) in dynamics, which could, in theory, be detected. This type of transition could be more likely in focal epilepsies (Mormann et al., 2007).

For prediction of seizures it is important to know the source of the EEG to be analyzed, and therefore intracranial recording is preferable over scalp EEG. A higher SNR is, of course, another advantage of intracranial recording. Still, quite a few works on seizure prediction using scalp EEG signals have been reported (see Table 1 in Mormann et al., 2007). As far as seizure prediction is concerned, a nonlinear dynamical

system theoretical analysis of both scalp and depth EEG is very popular. According to this form of analysis, the brain is considered as a nonlinear dynamical system, whose output is EEG. Since nonlinear dynamics is a well-developed theory, it can readily be used as a tool to analyze EEG. Different similarity measures among the EEG signals are examples of such tools. Navarro and coworkers identified a similarity measure that can be useful for seizure prediction in both scalp and depth EEG (Navarro et al., 2002). Also, this was one of the few seizure prediction methods applied on scalp EEG (most others were tested on depth EEG only). This measure is suitable for prolonged EEG recording, from several hours to days on end, of the occurrence of seizures. It is described as follows (Van Quyen et al., 1999).

1. A reference EEG segment S_{ref} (usually several minutes long) is chosen quite distant in time from any seizure. This is done independently in all EEG channels.
2. Let the time of the nth threshold crossing (set here to the zero of the signal) for a signal be denoted as T_n. Let $I_n = T_{n+1} - T_n$ be the time interval between two successive zero crossings. Form the *delay vector* $A_n = (I_n, I_{n-1}, \ldots, I_{n-m+1})$, where m is called the *embedding dimension* of the vector space consisting of A_n for different n values in the signal from the same channel. In Van Quyen et al. (1999) the value of m was taken to be 16.
3. A number of n's (say d) need to identified so that as many reference segments A_n (each with dimension m) from the signal can be selected to form a $d \times m$ matrix **B**. Each row of **B** is a segment of the EEG signal from a particular channel and m is the embedding dimension of each segment. Perform the *singular value decomposition* (SVD) on **B** (see below for more details on SVD). The component associated with the highest singular value is taken as the one representing the actual dynamics of the original EEG signal most appropriately (Van Quyen et al., 1999). Let this component be denoted by C_i, when it is coming from the ith EEG channel.
4. C_i derived above will have to be compared with C_j derived from a reference segment of the jth EEG channel. To compare the EEG signals from ith and jth channels more segments from each of them should be considered. The kth segment in the ith channel will give rise to C_{ik} as the snapshot of the most representative dynamics of the kth segment and similarly C_{jl} for the lth segment of the jth channel.
5. A distance measure between collection of all C_{ik}'s and collection of all C_{jl}'s $d(\{C_{ik}\}_k, \{C_{jl}\}_l)$ will give a similarity measure between the two

EEG channels. This similarity measure has been taken to be a *cross-correlation integral* in Van Quyen et al. (1999) as

$$d\left(\{C_{ik}\}_k, \{C_{jl}\}_l\right) = \frac{1}{MN}\sum_{k=1}^{M}\sum_{l=1}^{N}\Theta\left(\|C_{ik} - C_{jl}\| - r\right), \tag{6.13}$$

where
- Θ is the Heaviside function
- M is the cardinality of $\{C_{ik}\}_k$
- N is the cardinality of $\{C_{jl}\}_l$
- $\|\|$ is the Euclidean norm
- r is a predetermined distance, so that a fraction of the pairs (say 30%) C_{ik} and C_{jl} will be closer to each other by a distance less than r

The similarity measure (6.13) has been further refined to give another (and final) normalized similarity measure whose value lies between 0 and 1. The refined measure is defined as (Van Quyen et al., 1999)

$$D\left(\{C_{ik}\}_k, \{C_{jl}\}_l\right) = \frac{d\left(\{C_{ik}\}_k, \{C_{jl}\}_l\right)}{\sqrt{d\left(\{C_{ik}\}_k, \{C_{ik}\}_k\right) d\left(\{C_{jl}\}_l, \{C_{jl}\}_l\right)}}. \tag{6.14}$$

Let μ and σ be the mean and standard deviation of variations of the similarity measure D during a baseline period (say, for example, 250 s at the start of the recording) in interictal EEG recording. We denote here the statistical significance Σ of the deviation by the ratio

$$\Sigma = \frac{\gamma - \mu}{\sigma}, \tag{6.15}$$

whose p-value is given by the Chebyshev's inequality (for any statistical distribution of γ): $P(|\Sigma| \geq k) \leq 1/k^2$ where k is the chosen statistical threshold (Van Quyen et al. 1999) (Figure 6.16).

Before moving further ahead, let us briefly describe what singular value decomposition is. SVD is similar to PCA discussed in Section 2.2.1. However, SVD is a completely linear algebraic technique, without any statistical consideration. It is applicable to any $m \times n$ matrix, where $m \neq n$. For $m = n$, it reduces to diagonalization. It says that a rectangular matrix A_{mn} can be expressed as a product of three matrices $A_{mn} = U_{mm} \times S_{mn} \times V_{nn}^T$ (Baker, 2013), where $U^T U = I_m$, $V^T V = I_n$, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order.

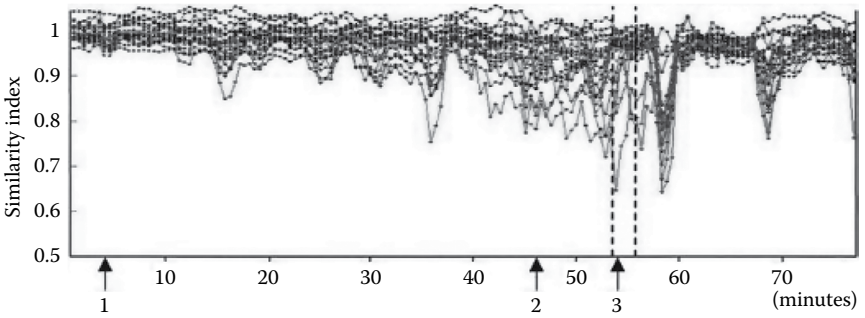


FIGURE 6.16
The time course of the similarity index for 21 scalp EEG channels of an epileptic patient showed progressive deviations until the starting of seizure from the point labeled 1. Label 3 is soon after the seizure onset. Dashed vertical lines indicate seizure onset and offset time. (Adapted from Navarro, V. et al., *Brain*, 125(3), 640, 2002.)

The columns of U are orthonormal eigenvectors of AA^T and the columns of V are orthonormal eigenvectors of A^TA. For a short and lucid introduction and some computational examples, see the tutorial by Baker (2013). The highest singular value of A is the component associated with the highest eigenvalue of V. SVD has been implemented in MATLAB® (type 'help svd' in the command prompt).

Applying the above-described similarity measure on scalp EEG recording, prediction of a seizure as much as 15 minutes ahead of onset was reported for one patient (Navarro et al. 2002). In this application, the value of the embedding dimension m was taken as 4. Traditional analyses of EEG signals have failed to detect specific changes preceding seizures by visual inspection of the recordings. The similarity index quantified the changes in dynamics relative to a constant reference window at the beginning of a preictal recording. Decreased similarity was observed before seizures in both intracranial and scalp EEG recordings (Mormann et al., 2007).

6.5 ROC Curve Analysis

Receiver operating characteristic (ROC) curve analysis is a very standard tool to ascertain the reliability of a binary classifier. Be it seizure detection or prediction, it is a binary classification problem, because the question is whether the seizure has been detected (successfully predicted) or not. Here we will describe ROC curve analysis for automatic seizure detection only. The same will hold for seizure prediction also.

As we have seen in Section 6.1, seizure detection has two parts, namely, detection of actual seizures and avoiding detection of seizure-like artifacts as seizure. We define the following four terms, which describe the whole situation accurately.

1. True positive (T_p) = Number of seizures detected correctly.
2. True negative (T_n) = Number of artifacts (i.e., nonseizures) identified as artifacts (i.e., nonseizures).
3. False positive (F_p) = Number of artifacts (i.e., nonseizures) detected as seizures.
4. False negative (F_n) = Number of seizures identified as artifacts and therefore not detected.

With the help of these four terms a *confusion matrix* can be created as in Table 6.3.

Next we define $\alpha = \dfrac{F_p}{T_n + F_p}$ and $\beta = \dfrac{F_n}{T_p + F_n}$. Then we define *sensitivity* $= 1 - \alpha = \dfrac{T_n}{T_n + F_p}$ and *sensitivity* $= 1 - \beta = \dfrac{T_p}{T_p + F_n}$. The ROC curve is the plot of $(\alpha, 1-\beta)$, that is, 1—specificity vs. sensitivity. Or, in other words, the ROC curve is a plot of probability of true positive $(T_p/(T_p + F_n))$ or the *benefit* on the Y-axis vs. probability of false positive $(F_p/(T_n + F_p))$ or the *cost* on the X-axis. The area under the (ROC) curve (AUC), which lies between 0 and 1, gives a performance measure of seizure detection (also of seizure prediction, as the case may be) algorithm (Bradley, 1997). Figure 6.17 presents an example of ROC curve analysis for performance evaluation of an automatic seizure detection algorithm in depth EEG recording of 14 patients with epilepsy. The data was obtained from the Freiburg Seizure Prediction Project, University of Freiburg, Freiburg, Germany.

Here, one important thing to note is that the ROC curve computation in Figure 6.17 has not been done in the conventional way. A typical conventional approach is to take a threshold parameter (say, merely for the sake of example, signal amplitude), so that above the threshold the signal will be

TABLE 6.3

Confusion Matrix

	Predicted Class		
True Class	**–ve**	**+ve**	
–ve	T_n	F_p	C_n
+ve	F_n	T_p	C_p
	R_n	R_p	N

Note: $C_n = T_n + F_p$, $R_n = T_n + F_n$, etc. $N = R_n + R_p = C_n + C_p$.

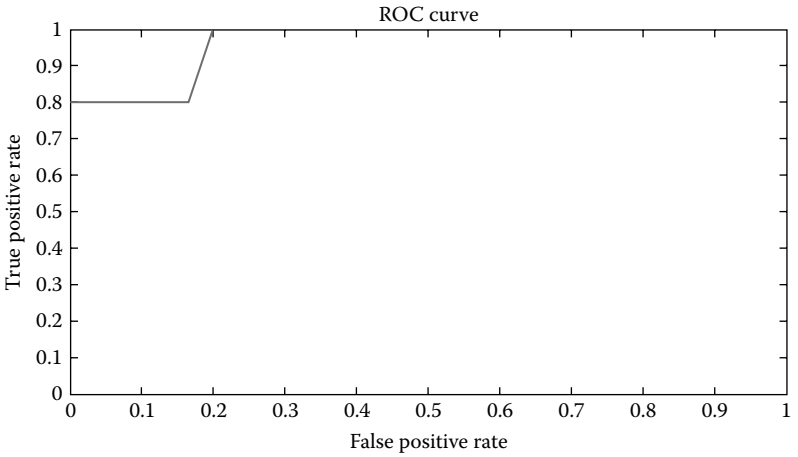


FIGURE 6.17
ROC curve for detection of seizures of 14 patients. A total of 51 seizures were recorded out of which 48 were automatically detected as shown in Table 6.4. Area under the curve is 0.959. (Taken from Malali, A. et al., *Biomed. Signal Proc. Control*, 26, 98, 2016.)

considered as the seizure signal and below the threshold the signal will be considered as the background signal. Now, this threshold value will have to be varied within an interval. If we take 50 values from within that interval with equal distance between any two successive values, then we get 50 different (cost, benefit) points inside the unit square, one for each threshold value. The ROC curve is the curve that traverses through (0, 0), all those 50 points, and (1, 1). Since in seizure detection or prediction the threshold will always have to be patient specific, for each patient in a group study there will have to be one ROC curve. The group ROC curve may be taken as the average of all the individual ROC curves. But in Figure 6.17, we have shown an alternative way to calculate the group ROC curve, which is much easier to compute. Nowadays, a study of seizures almost always involves a large group of patients. The average AUC across all the patients can be used as a performance measure for the detection (prediction) algorithm's performance.

Table 6.4 gives the complete list of detected seizures by an algorithm called the *differential variance slope spectrum* (DVSS) algorithm (Malali et al., 2016) for each of the 14 patients whose seizures were tested for automatic detection. Average detection delay across all the detected seizures was 5.02 s. False positives are reported after running DVSS for 24–26 hours of interictal recording for each patient except Patient 2, for whom interictal data was not available. However, the average false positive rate across other 13 patients was close to 1 per day. Therefore, 1 was assigned for Patient 2 as the number of false positives for 24 hours of interictal recording. It was also necessary for drawing the ROC curve as shown in Figure 6.17. Here the AUC is 0.959, indicating a good performance of the detection algorithm.

TABLE 6.4

Patients Detail

Patient No.	No. of Seizures Detected/No. of Seizures Recorded for the Patient	False Positives: No. of False Detections in 24 Hours
2	3/3	1
3	4/5	2
4	4/4	1
6	3/3	1
7	2/2	1
8	2/2	2
10	3/3	1
12	4/4	0
13	2/2	0
14	3/4	0
15	4/4	2
16	4/5	1
17	5/5	1
20	5/5	3
Total no. of seizures	48/51	

Source: Taken from Malali, A. et al., *Biomed. Signal Proc. Control*, 26, 98, 2016.

References

Adeli, H., Z. Zhou, and N. Dadmehr, Analysis of EEG records in an epileptic patient using wavelet transform, *J. Neurosci. Methods*, **123**: 69–87, 2003.

Baker, K., Singular value decomposition tutorial, 2005, 2013, available at http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf (accessed August 9, 2014).

Ball, T., M. Kern, I. Mutschler, A. Aertsen, and A. Schulze-Bonhage, Signal quality of simultaneously recorded invasive and non-invasive EEG, *NeuroImage*, **46(3)**: 708–716, 2009.

Blanke, O., G. Lantz, M. Seeck, L. Spinelli, R. Grave de Peralta, G. Thut, T. Landis, and C. M. Michel, Temporal and spatial determination of EEG-seizure onset in the frequency domain, *Clin. Neurophysiol.*, **111**: 763–772, 2000.

Bonelli, S. B., S. Lurger, F. Zimprich, E. Stongmann, E. Aseem-Hilger, and C. Baumgartner, Clinical seizure lateralization in frontal lobe epilepsy, *Epilepsia*, **48(3)**: 517–523, 2007.

Bradley, A. P., The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recogn.*, **30(7)**: 1145–1159, 1997.

Cecchin, T., R. Ranta, L. Koessler, O. Caspery, H. Vespignani, and L. Millard, Seizure lateralization in scalp EEG using Hjorth parameters, *Clin. Neurophysiol.*, **121**: 290–300, 2010.

Fisher, R. S., W. van Emde Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel, Jr., Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE), *Epilepsia*, **46(4)**: 470–472, 2005.

Gotman, J. and P. Gloor, Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG, *Electroencephal. Clin. Neurophysiol.*, **41**: 513–529, 1976.

Guo, L., D. Rivero, J. Dorado, J. R. Rabunal, and A. Pazos, Automatic seizure detection in EEGs based on line length feature and artificial neural networks, *J. Neurosci. Methods*, **191**: 101–109, 2010.

Holmes, M. D., D. M. Tucker, J. M. Quiring, S. Hakimian, J. W. Miller, and J. G. Ojemann, Comparing noninvasive dense array and intracranial electroencephalography for localization of seizures, *Neurosurgery*, **66(2)**: 354–362, 2010.

Khan, Y. U. and J. Gotman, Wavelet based automatic seizure detection in intracerebral electroencephalogram, *Clin. Neurophysiol.*, **114**: 898–908, 2003.

Logesparan, L., A. J. Casson, and E. Rodriguez-Villegas, Optimal features for online seizure detection, *Med. Biol. Eng. Comput.*, **50**: 659–669, 2012.

Lopes da Silva, F., W. Blanes, S. Kalitzin, J. Parra, P. Suffczynski, and D. N. Velis, Epilepsies as a dynamical disease of brain systems: Basic models of the transition between normal and epileptic activity, *Epilepsia*, **44**: 72–83, 2003.

Majumdar, K., Differential operator in seizure detection, *Comput. Biol. Med.*, **42(1)**: 70–74, 2012.

Majumdar, K. and P. Vardhan, Efficacy of differential operators in brain electrophysiological signal processing: A case study in epilepsy, *IEEE Trans. Neural Syst. Rehabil. Eng.*, **19(4)**: 356–365, 2011.

Malali, A., G. Chaitanya, S. Gowda, and K. Majumdar, Analysis of cortical rhythms in intracranial EEG by temporal difference operators during epileptic seizures, *Biomed. Signal Proc. Control*, **26**: 98–108, 2016.

Marks, W. J. and K. D. Laxer, Semiology of temporal lobe seizures: Value in lateralizing seizure focus, *Epilepsia*, **39(7)**: 721–726, 1998.

Mormann, F., A. G. Andrzejak, C. E. Elger, and K. Lehnertz, Seizure prediction: The long and winding road, *Brain*, **130(2)**: 314–333, 2007.

Murro, A. M., Y. D. Park, D. W. King, B. B. Gallagher, J. R. Smith, K. J. Meador, and W. Littleton, Localization of temporal lobe seizures with quantitative EEG, *Electroencephal. Clin. Neurophysiol.*, **86**: 88–93, 1993.

Navarro, V., J. Martinerie, M. Le van Quyen, S. Celmenceau, C. Adam, M. Baulac, and F. Varela, Seizure anticipation in human neocortical partial epilepsy, *Brain*, **125(3)**: 640–655, 2002.

Noachtar, S. and A. S. Peters, Semiology of epileptic seizures, *Epilepsy Behav.*, **15(1)**: 2–9, 2009.

Pathak, A., A. Ramesh, A. Mitra, and K. Majumdar, Automatic seizure detection by modified line length and Mahalanobis distance function, *Biomed. Signal Proc. Control*, (accepted), 2017.

Pillai, J. and M. R. Sperling, Interictal EEG and diagnosis of epilepsy, *Epilepsia*, **47(Suppl. 1)**: 14–22, 2006.

Qu, H. and J. Gotman, A patient-specific algorithm for detection of seizure onset in long-term EEG monitoring: Possible use as a warning device, *IEEE Trans. Biomed. Eng.*, **44(2)**: 115–122, 1997.

Rana, M., Integrate epilepsy treatment into PHCs, *Deccan Herald*, **68(318)**: 11, 2015.

So, E. L., Interictal epileptiform discharges in persons without a history of epilepsy: What do they mean? *J. Clin. Neurophysiol.*, **27(4)**: 229–238, 2010.

Temuçin, C. M., A. B. Tokçaer, and E. Bilir, Detection of EEG background abnormalities in epilepsy by a new spectral index, *Clin. Neurophysiol.*, **116**: 933–947, 2005.

van Putten, M. J. A. M., The revised brain symmetry index, *Clin. Neurophysiol.*, **118**: 2362–2367, 2007.

van Putten, M. J. A. M., T. Kind, F. Visser, and V. Lagerburg, Detecting temporal lobe seizures from scalp EEG recordings: A comparison of various features, *Clin. Neurophysiol.*, **116**: 2480–2489, 2005.

Van Quyen, M. L., J. Martinerie, M. Boulac, and F. Varela, Anticipating epileptic seizures in real time by a non-linear analysis of similarity between EEG recordings, *NeuroReport*, **10**: 2149–2155, 1999.

7

Brain–Computer Interface

In a brain–computer interface or BCI the brain provides input directly to the computer through brain signals, without the aid of a mechanically operated device like a keyboard. Because of the ease of collection, high portability of the acquisition systems, low cost, and minimum risk involved, EEG has traditionally been the most preferred brain signal for this purpose. Roughly speaking, BCI is based on the classification of EEG signals. A computer is 'trained' to recognize the particular EEG pattern from a particular individual's brain during the performance (or the desire for performance) of a specific task. So, BCI implementations are largely EEG pattern–recognition tasks.

One immediate benefit of the BCI technique is better rehabilitation of persons with severe neuromuscular disabilities, such as amyotrophic lateral sclerosis, brainstem stroke, and spinal cord injury (Wolpaw et al., 2002). The intent of the user is determined from a wide range of EEG signals, such as event-related potentials, and mu and beta rhythms, recorded from the scalp. However, as we have seen in Section 3.1, the quality of scalp recording is affected by tissue impedances of the skull and other layers in the head. Also, scalp EEG is more prone to contamination by noise and artifacts compared to depth EEG. Therefore, significantly increasing the SNR of scalp EEG before actual processing for BCI applications is an essential prerequisite. Wherever possible, selecting appropriate channels with a high SNR is another alternative.

The purpose of a BCI is to interpret user intentions by means of monitoring cerebral activity. Brain signals involve numerous simultaneous phenomena related to cognitive tasks. Most of them are still incomprehensible, and their origins are unknown. However, the physiological phenomena of some brain signals have been decoded in such ways that people may learn to modulate them at will, to enable the BCI systems to interpret their intentions. These signals are regarded as possible control signals in BCIs.

Numerous studies have described a vast group of brain signals that might serve as control signals in BCI systems. The control signals employed in current BCI systems are visually evoked potentials (VEPs), slow cortical potentials (SCPs), P300-evoked potentials, sensorimotor rhythms (Nicolas-Alonso and Gomez-Gil, 2012), and movement-related potentials (MRPs) (Bashashati et al., 2007) (Table 7.1).

TABLE 7.1

Summary of Control Signals Useful for BCI Applications

Signal	Physiological Phenomena	Number of Choices	Training	Information Transfer Rate (bits/min)
VEP	Brain signal modulations in the visual cortex	High	No	60–100
SCP	Slow voltage shifts in the brain signals	Low (2 or 4, very difficult)	Yes	5–12
P300	Positive peaks due to infrequent stimulus	High	No	20–25
Sensorimotor rhythms	Modulations in sensorimotor rhythms synchronized to motor activities	Low (2, 3, 4, 5)	Yes	3–35

Source: Taken from Nicolas-Alonso, L.F. and Gomez-Gil, J., *Sensors*, 12, 1211, 2012.

VEPs are brain activity modulations that occur in the visual cortex after receiving a visual stimulus. These modulations are relatively easy to detect since the amplitude of VEPs increases enormously as the stimulus is moved closer to the central visual field.

SCPs are slow voltage shifts in the EEG that last from a second to several seconds. SCPs belong to the part of the EEG signals below 1 Hz. SCPs are associated with changes in the level of cortical activity. Negative SCPs correlate with increased neuronal activity, whereas positive SCPs coincide with decreased activity in individual cells. These brain signals can be self-regulated by both healthy users and paralyzed patients to control external devices by means of a BCI. SCP shifts can be used to move a cursor and select the targets presented on a computer screen (Nicolas-Alonso and Gomez-Gil, 2012).

P300 was discussed in Chapter 4. The more novel or infrequent the stimulus, the larger the peak amplitude. But the amplitude comes down as the brain gets habituated to the stimulus. Since P300 is the most prominent among all the ERPs and with the largest displacement from the baseline, its detection is relatively easy and the automatic detector does not need any training (simple thresholding will do).

Sensory motor rhythms are mu and beta rhythms (Chapter 1). The amplitude of the sensorimotor rhythms varies due to the cerebral activity related to motor tasks with or without the actual movements. Motor rhythms are produced as a result of mental rehearsal of a motor act without any overt motor output. These rhythms are natural candidates to control BCIs. Sensorimotor rhythms can endure two kinds of amplitude modulations known as event-related desynchronization (ERD) and event-related synchronization (ERS) that are generated by sensory stimulation, motor behavior, and mental

imagery. ERD involves amplitude suppression (because of desynchronous firing of neurons) of the rhythm, and ERS implies amplitude enhancement (because of synchronous firing of neurons). We will come back to ERD and ERS again in this chapter.

MRPs represent averaged electroencephalographic activity before and after a voluntary movement. They begin with a slowly rising negativity and progress to a steeper, later negativity starting about 400 ms before the onset of movement. They are followed by the motor potential, which is seen partly before and partly after the movement. The initial slope of motor potential occurs just before the onset of electromyographic (EMG) activity, is focal topographically over the primary motor cortex, and probably represents activation of the primary motor cortex. This contralateral focal negativity persists for 30–50 ms after the onset of EMG activity. It then drops off in the central and parietal regions. Subsequently, the peak negativity shifts toward the anterior contralateral area, where it reaches the highest negativity of the recording (Hallett, 1994).

7.1 Preprocessing and Signal Enhancement

All the standard techniques, namely, FIR and IIR filtering, PCA, and ICA that we discussed in Chapter 2 are commonly used for preprocessing EEG before being used for the BCI. Since the EEG signal on the scalp is very faint (of the order of tens of microvolts only), even after removal of artifacts, enhancing the signal is particularly useful for a successful application. We have seen in Chapter 6 how useful differential operators can be in enhancing the signal during an epileptic seizure relative to the background. Here we will discuss a similar technique, not for single EEG channels, but for two-dimensional scalp potential. This is called the surface Laplacian (SL) and has been used widely for signal enhancement in BCI applications (see Figure 2 of Bashashati et al., 2007).

Brain electric signals recorded on the scalp are distorted by tissue impedance due to the skull and tissue layers in the head. This distortion acts as a weak, low-pass filter, which causes the potentials at the scalp to appear blurred. They may also become contaminated by unwanted interferences from the reference electrode. Correction of the distortion of scalp potentials will have to take into account the surface geometry of the tissue layers. Nevertheless, second-order spatial derivative–based SL, which acts as a high-pass filter to undo the low-pass effect of blurring, can be applied to sharpen the spatial frequency resolution and to eliminate common activities at all electrodes (the so-called referential artifacts).

Let us consider the following simplified example. In many literature reports, the human head is modeled as concentric spheres (see Figure 3.1).

Let us further simplify the human head model as one single sphere. If R is the radius of the spherical scalp surface, the equation of the scalp surface can be written as

$$\left.\begin{aligned} x &= R\cos\theta\sin\phi = f(\theta,\phi) \\ y &= R\cos\theta\cos\phi = g(\theta,\phi) \\ z &= R\sin\theta = h(\theta,\phi) \end{aligned}\right\}, \tag{7.1}$$

where θ, ϕ are parameters. Equation 7.1 is the parametric representation of the scalp surface on which the potential or EEG needs to be measured. The conventional Laplacian $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is calculated on the Euclidean XY plane. For the spherical scalp potential ∇^2 will have to be approximated on the parametric surface (7.1). By this is meant the SL. The electrode at the location (x,y,z) is normal to the scalp surface at that point. So, it records the electric field normal to the scalp surface at that point. SL at (x,y,z) exactly gives the electric field normal to the scalp surface at that point. In other words, the electrode at (x,y,z) does not record the raw potential at that point, but the SL of the raw potential $V(x,y,z)$ at (x,y,z).

Let the scalp potential at (x,y,z) be $V(x,y,z) = V(f(\theta,\phi), g(\theta,\phi), h(\theta,\phi))$.

With the above overview of EEG-based BCI control signals and their important characteristics, we need to perform suitable feature extraction, which will lead to task-specific classification so that the computer can associate the signals with the tasks to be performed. Among the above BCI control signals SCPs, mu and beta rhythms should best be represented in the frequency domain.

7.2 Frequency Domain Features

Let us start with a case study of cursor movement by transmitting the intention of the brain through scalp EEG to a computer (Figure 7.1). 8–12 Hz mu rhythms and 13–28 Hz beta rhythms in the scalp EEG are produced in sensory motor cortex and associated areas (the time series signal in Figure 7.1a). They are directly related to movements or movement intentions. With practice, by subjects amplitudes of mu and beta rhythms can be enhanced or suppressed so that the frequency domain EEG signal bearing these signatures can be given as input commands to a computer. With training, the computer can associate specific signatures (illustrated in Figure 7.2) with specific cursor movements.

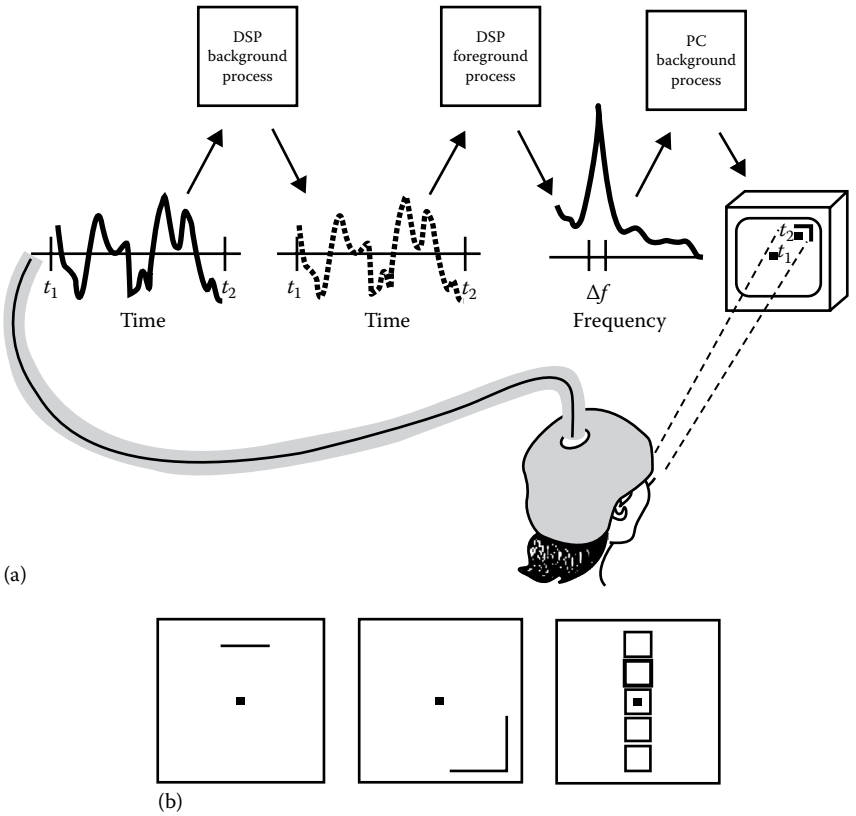


FIGURE 7.1
Block diagram of a cursor-movement task as a BCI performance. For want of space the task has been shown for only one EEG signal (channel). (a) BCI operation. The scalp voltage is amplified, digitized, spatially filtered, and frequency analyzed 10 times/s. The amplitude in a specific frequency band is translated into a cursor movement. This is performed by foreground and background processes on the digital signal processing (DSP) board and the PC. (b) Three different control modes. On the left is the basic one-dimensional mode in which the target is on the top or the bottom edge and the cursor, which begins in the middle and moves vertically controlled by the EEG until it reaches the top or bottom edge. In the middle is the two-dimensional mode, in which the target is at one of four or more positions on the periphery of the screen, and the cursor moves both vertically and horizontally controlled by the EEG until it reaches the periphery. On the right is the graded one-dimensional mode, in which the target is in the highlighted box of a series of boxes arranged vertically on the screen and the cursor begins in the middle and moves vertically controlled by the EEG until it stays in one box for a defined period (e.g., 1 s) and thereby selects it. (*Continued*)

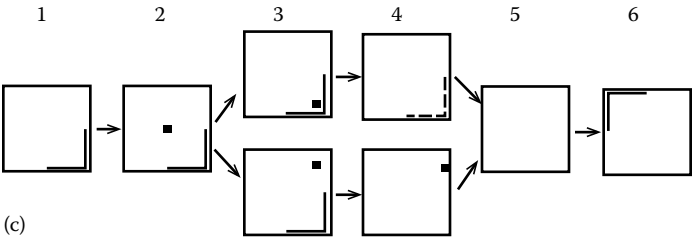


FIGURE 7.1 (*Continued*)
Block diagram of a cursor-movement task as a BCI performance. For want of space the task has been shown for only one EEG signal (channel). (c) Sequence of events during a trial. (1) The trial begins when a target appears in one corner. (2) After a brief period (e.g., 1 s) that allows the subject to see the location of the target and initiate the proper EEG, the cursor appears in the center. (3) The cursor moves, controlled by the EEG, until it reaches the periphery. (4) If it reaches the part occupied by the target, a hit is registered, the cursor disappears, and the target flashes for 1 s as a reward. If it reaches another part, a miss is registered, the target disappears, and the cursor remains fixed on the screen for 1 s. (5) The screen is blank for 1 s. (6) The next target appears. (Taken from Wolpaw, J.R. et al., *IEEE Trans. Rehabil. Eng.*, 8(2), 222, 2000.)

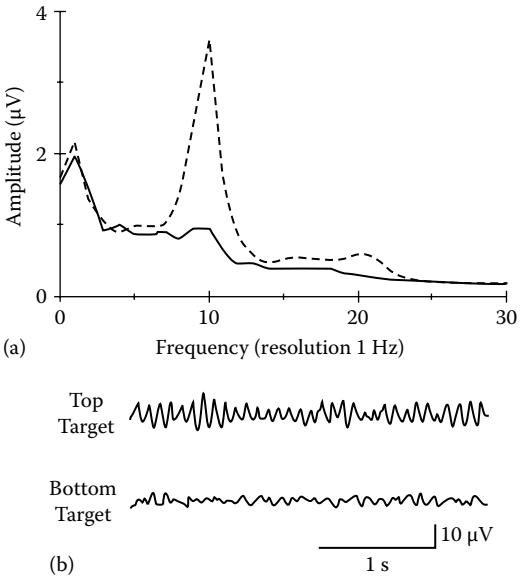


FIGURE 7.2
(a) Frequency spectra of EEG recorded over sensorimotor cortex of a trained subject when the target is at the bottom (solid) or at the top (dashed) of the video screen. The main difference between the two spectra is in the 8–12 Hz mu rhythm band (and, to a lesser extent, in an 18–23 Hz beta rhythm band). Differences at other frequencies are absent or minimal. (b) Sample EEG traces accompanying top or bottom targets. The mu rhythm is prominent with the top target, and minimal with the bottom target. (Taken from Wolpaw, J.R. et al., *IEEE Trans. Rehabil. Eng.*, 8(2), 222, 2000.)

Here, signals are extracted from the EEG controlled movement of a cursor on a video screen (Figure 7.2). Two factors determine the accuracy and speed of cursor movement, namely, the skill of the subject, that is, the magnitude and consistency of the control signals that the subject produces and the signal-to-noise ratio achieved by the online analysis that extracts these control signals from the EEG and translates them into cursor movement. The noise has two sources: non-EEG artifacts including EMG, EKG, and eye-movement and eye-blink potentials; and non-mu EEG components, such as the visual alpha rhythm (McFarland et al., 1997). Here, the signal signature collection is performed through a spatial filtering scheme.

First, the montage is changed to the common average reference (CAR) montage, in which the average of all channels is subtracted from each of the channels. CAR provides a nearly reference-free EEG recording (McFarland et al., 1997). It accentuates components with highly localized distributions (such as task-specific mu and beta activities). Also, it removes events present across the channels. In other words, it accentuates localized activities and removes more diffused activities, which is characteristic of a high-pass (spatial) filter.

Next, we calculate the Laplacian across the electrodes (spatial Laplacian). For each electrode location it calculates the second derivative of the instantaneous spatial voltage distribution, thereby emphasizing activities originating in the radial sources immediately below the electrode (McFarland et al., 1997). Thus, it also works like a high-pass spatial filter that accentuates localized activities and suppresses more diffuse or spread-out activities. High spatial resolution can be achieved by using dense array electrodes (more than 200) spread over the entire scalp. The value of the Laplacian at each electrode location is calculated by combining the value at that location with the values of a set of surrounding electrodes. The distances to the set of surrounding electrodes determine the spatial filtering characteristics of the Laplacian. As the distance decreases the Laplacian becomes more sensitive to potentials with higher spatial frequencies and less sensitive to those with lower spatial frequencies.

The Laplacian should be calculated in two different ways with two different sets of surrounding electrodes, namely, nearest neighbor (adjacent) electrodes (called the small Laplacian) and next-nearest neighbor electrodes (called the large Laplacian). In Figure 7.3, the arrangement of neighboring electrodes for the calculation of the small Laplacian and the large Laplacian has been shown with a typical example.

The Laplacian is calculated as

$$V_i^{Lap} = V_i^{ER} - \sum_{j \in S_i} g_{ij} V_j^{ER}, \tag{7.2}$$

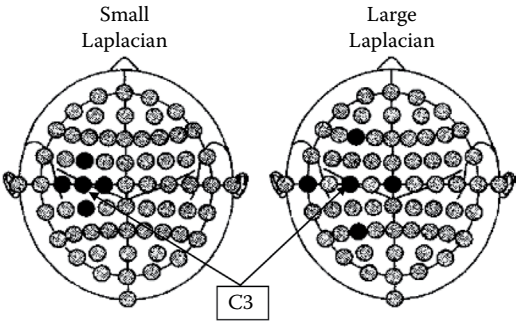


FIGURE 7.3
The arrangement of the neighboring electrodes of C3 has been shown for the small Laplacian (left) and the large Laplacian (right) calculations. (Adapted from McFarland, D.J. et al., *Electroencepahal. Clin. Neurophysiol*., 103, 386, 1997.)

where $g_{ij} = \dfrac{1/d_{ij}}{\sum_{j \in S_i} 1/d_{ij}}$. d_{ij} is the distance between electrodes i, and j, S_i is the set of electrodes surrounding the electrode i. V_i^{Lap} stands for the Laplacian calculated at the electrode i, and V_i^{ER} is the potential calculated between the electrode i and the reference electrode. The small Laplacian and the large Laplacian are calculated by varying the sets S_i, which are illustrated in Figure 7.3. For small Laplacian calculation at C_3, we need to consider the neighborhood of C_3 consisting of the black dotted electrodes in the left panel of Figure 7.3 (the immediate neighbors of C_3 and C_3 itself). For large Laplacian calculations we need to consider the black dotted electrodes in the right panel of Figure 7.3.

Please note that the Laplacian involves spatial double derivatives. However, McFarland and colleagues described the Laplacian only as a single difference operation according to Equation 7.2 (McFarland et al., 1997), which may be misleading. Nevertheless, it gave the desired results as reported in McFarland et al. (1997) and Wolpaw et al. (2000).

The waveforms resulting from each of the spatial filtering methods described above is then subjected to an autoregressive spectral analysis known as the maximum entropy method (MEM) (McFarland et al., 1997). Now, the challenge is to build a representative waveform that will represent the brain cognition while performing a particular task from the trial waveform of that task during the training session. MEM is a well-established method for accomplishing this from one single trial.

It is a method for the power spectrum estimation of a signal, like the FFT-based power spectrum estimation. MEM is another way to implement autoregressive (AR) power spectrum estimation (see Chapter 14 of Proakis and Manolakis, 2007). In an AR model of order p, the signal value $x(p)$ is

estimated from the given values $x(p-1)$, $x(p-2)$, $x(p-3)$, …, and $x(1)$. In the MEM, the power spectrum $P(f)$ associated with frequency f is given by

$$P(f) = \frac{\hat{E}_p}{\left|1 + \sum_{k=1}^{p} \hat{a}_p(k) e^{-i2\pi fk}\right|^2}, \tag{7.3}$$

where $i = \sqrt{-1}$, $\hat{E}_p$ is the estimated least square error and $\hat{a}_p(k)$ is the estimated kth parameter of the AR model of order p (Equation 14.3.20 in Proakis and Manolakis (2007); also see Makhoul (1975) for a very lucid description of the AR model). The parameters are estimated by the following equation:

$$\begin{bmatrix} \gamma_{xx}(0) & \gamma_{xx}(-1) & \cdots & \gamma_{xx}(-p+1) \\ \gamma_{xx}(1) & \gamma_{xx}(0) & \cdots & \gamma_{xx}(-p+2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{xx}(p-1) & \gamma_{xx}(p-2) & \cdots & \gamma_{xx}(0) \end{bmatrix} \begin{bmatrix} \hat{a}_p(1) \\ \hat{a}_p(2) \\ \vdots \\ \hat{a}_p(p) \end{bmatrix} = \begin{bmatrix} \gamma_{xx}(1) \\ \gamma_{xx}(2) \\ \vdots \\ \gamma_{xx}(p) \end{bmatrix}, \tag{7.4}$$

where $\gamma_{xx}(i)$ is the autocorrelation of x, which can be expressed in the biased form as

$$\gamma_{xx}(i) = \frac{1}{N} \sum_{n=0}^{N-i-1} x(n) x(n+i), \tag{7.5}$$

N is the number of sample points in x. Estimating $\hat{E}_p$ in (7.3) is tricky. It can be written as $\hat{E}_p = \sigma^2$, where σ^2 is the variance of a Gaussian random noise signal (Kay, 2010), given by

$$\sigma^2 = \gamma_{xx}(0) + \sum_{k=1}^{p} \hat{a}_p(k) \gamma_{xx}(-k), \tag{7.6}$$

as in equation (14.3.8) of Proakis and Manolakis (2007).

The most prominent features of EEG signals for the BCI have been ERS and ERD, both of which we have mentioned earlier in this chapter. When the brain is involved in certain tasks, the signal power goes up or down in certain frequency bands at certain spatial locations of the EEG channels. When the signal power goes up it happens due to synchronous or simultaneous firing of a large number of neurons, which is termed as event-related synchronization (ERS). Similarly, when the signal power goes down it happens due to the absence of synchronous firing of the neurons, which is termed as event-related desynchronization (ERD). A measure of the same has been defined in the following way (Pfurtscheller and Lopes da Silva, 1999).

To obtain percentage values for ERD/ERS, the power within the frequency band of interest in the period after the event is given by A whereas that of the preceding baseline or the reference period is given by R. ERD or ERS is defined as the percentage of power decrease or increase, respectively, according to the expression

$$ERD\% = \frac{A-R}{R} \times 100. \tag{7.7}$$

Note that lesser the ERD the more will be the ERS and vice versa. We can think of ERD + ERS = 100%. If ERD is 40% then ERS is 60%.

7.3 Time Domain Features

P300 is the strongest response of our brain to a novel stimulus manifested in the EEG. An illuminated letter or an illuminated digit in a 6×6 matrix as shown in Figure 7.4 is one such novel stimulus. The resultant waveform of P300 in a subject will be slightly different from stimulus to stimulus, which is utilized to determine which letter or digit the brain is perceiving. In Figure 7.5 the P300 response of the letter P has been shown in the EEG channels Fz, Cz, and Pz recorded from two subjects.

The P300 waveform associated with each stimulus (letter) is quite faint. To improve the prominence of the waveform associated with each stimulus the averaged waveform across several trials needs to be used. The P300 waveform first averaged within single subjects and then averaged within each of two different classes of subjects across trials is shown in Figure 7.5. One of the classes consists of subjects on wheelchairs and the other consists of subjects

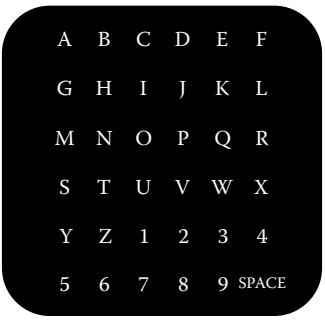


FIGURE 7.4
The stimulus matrix monitored by the subject. Every 125 ms a randomly chosen row or a randomly chosen column is intensified. (Taken from Donchin, E. et al., *IEEE Trans. Rehabil. Eng.*, 8(2), 174, 2000.)

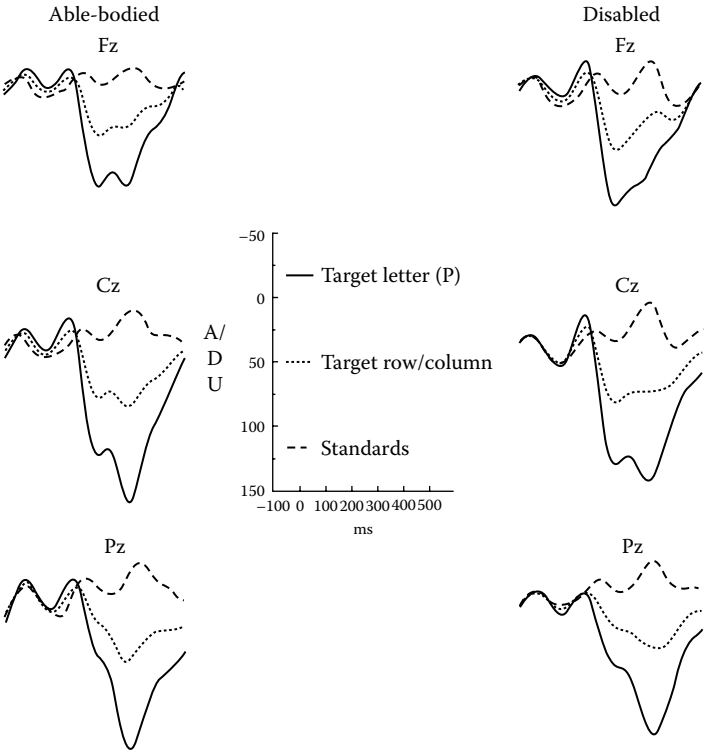


FIGURE 7.5
The P300 waveform averaged over two classes of subjects, namely, wheelchair bound and able-bodied during visualization of the letter P (solid line). P is indicated when the row and the column in the matrix in Figure 7.4 containing P are both lighted up. The averaged P300 associated with the row and the column has been shown as a dotted line. Positivity is in the downward direction. (Taken from Donchin, E. et al., *IEEE Trans. Rehabil. Eng.*, 8(2), 174, 2000.)

without motor impairment (the so-called able-bodied class). The P300 waveform shown in Figure 7.5 is associated with visualization of the letter P. The trial-averaged waveform associated with a particular stimulus from each subject is wavelet transformed using Daubechies 4 (DB4) wavelets. These transformed signals are then fed to the computer for moving the cursor to the particular character (stimulus) in the matrix shown in Figure 7.4.

7.4 Signal Analysis

The motor activity, both actual and imagined causes attenuation or augmentation in oscillatory activities in localized neuronal populations in the form of ERD or ERS, respectively. The *common spatial pattern* (CSP) algorithm is one

of the most efficient ways to calculate spatial filters for detecting ERD and ERS. In other words, it selects the channels with optimum ERD and ERS out of all the EEG channels, say N in number. CSP is described below on raw EEG data during single trial hand and foot movement following Wang et al. (2005).

Let X_F be a $N \times T_1$ matrix of EEG signals from N channels with T_1 sample points. These T_1 sample points cover the duration of foot movement. Similarly, let X_H be a $N \times T_2$ matrix, where T_2 is the number of sample points covering the entire duration of hand movement. The normalized spatial covariance of scalp EEG can be represented as

$$R_F = \frac{X_F X_F^T}{trace\left(X_F X_F^T\right)} \quad \text{and} \quad R_H = \frac{X_H X_H^T}{trace\left(X_H X_H^T\right)}, \tag{7.8}$$

where

M^T is the transpose of the matrix M

$trace(M)$ is the sum of diagonal elements of M

The averaged normalized covariance $\bar{R}_F$ and $\bar{R}_H$ are calculated taking the average of R_F and R_H across the foot-movement and hand-movement trials, respectively. The composite spatial covariance can be factored as

$$R = \bar{R}_F + \bar{R}_H = U_0 \Sigma U_0^T, \tag{7.9}$$

where

Σ is the $N \times N$ diagonal matrix of eigenvalues of R

U_0 is the $N \times N$ matrix whose rows are eigenvectors of R

Decomposing the $N \times N$ square matrix R as $R = U_0 \Sigma U_0^T$ is known as diagonalization or reduction into the canonical form and can be seen in any standard book of linear algebra. In (7.9) we need to rewrite Σ so that the diagonal elements are in decreasing order from upper-left to bottom-right corners. Also if the ith diagonal element of Σ is to be placed at the jth position in the diagonal due to the sorting operation, then the ith row of U_0 needs to be placed at the jth row position and ith column of U_0^T needs to be placed at the jth column position. For all subsequent calculations (7.9) will be treated as sorted accordingly. The whitening transformation matrix

$$P = \Sigma^{-\frac{1}{2}} U_0^T \tag{7.10}$$

transforms the average covariance matrices as

$$S_H = P \bar{R}_H P^T \quad \text{and} \quad S_F = P \bar{R}_F P^T. \tag{7.11}$$

S_H and S_F share common eigenvectors and the sum of corresponding eigenvalues of the two is always unity as the diagonalization operation on them yields,

$$S_H = U\Sigma_H U^T, \quad S_F = U\Sigma_F U^T, \quad \text{and} \quad \Sigma_H + \Sigma_F = I, \tag{7.12}$$

where I is the $N \times N$ identity matrix. It is clear from (7.12) that the eigenvectors of S_H with large eigenvalues will be eigenvectors of S_F with small eigenvalues and vice versa, because the sum of the two eigenvalues will always be 1. This is the key to having prominent discriminative features associated with foot movement and hand movement. The features that are strong during the foot movement will be weak during the hand movement and vice versa.

Now let us determine a combined set of features (for both the movements) by projection of the acquisitioned EEG signals during the movements. The projection matrix W is denoted as

$$W = U^T P, \tag{7.13}$$

where U is given by (7.12), and P is given by (7.10). With the projection matrix W given by (7.13), the original EEG matrix X ($X = X_F$ or $X = X_H$ as the case may be) for a single trial can be transformed as

$$Z = WX. \tag{7.14}$$

The columns of W^{-1} are known as common spatial patterns and these CSPs are the time-invariant EEG source distribution vectors (Ramoser et al., 2000) associated with particular tasks. Z is the projected EEG such that the separation of the variance of the signals (rows of Z) will be maximum among the tasks. By virtue of the sorting in (7.9), the first few rows (say m in number) of Z given by (7.14) will be strongly associated with one of the tasks (say hand movement). Then the last m rows will be weakly associated with the same task. The feature vectors are determined by the following quantities (Ramoser et al., 2000):

$$f_p = \log\left(\frac{\text{var}(Z_p)}{\sum_{i=1}^{2m} \text{var}(Z_i)}\right), \tag{7.15}$$

where Z_i is the ith row of Z. In (7.15) only the specified $2m$ rows of Z have been considered. Now consider the feature vector $\left(f_1^F, f_2^F, f_3^F, \ldots, f_{2m}^F\right)$ associated with $Z_F = WX_F$ and the $\left(f_1^H, f_2^H, f_3^H, \ldots, f_{2m}^H\right)$ associated with $Z_H = WX_H$ for different trials of hand movements and foot movements (it will be practical to keep the number of trials associated with the two different tasks almost

equal, else a biased training of the classifier may occur). There will be as many vectors $\left(f_1^F, f_2^F, f_3^F, \ldots, f_{2m}^F\right)$ as there are foot-movement trials and there will be as many vectors $\left(f_1^H, f_2^H, f_3^H, \ldots, f_{2m}^H\right)$ as there will be hand-movement trials. Taking this collection of trials (both types) as the training set we can design a classifier that will be able to distinguish between these two types of movements (or movement intentions) in a test set of a mixture of feature vectors (unmarked by suffix *F* or *H*) of those two types of trials. We will discuss about the classifiers in the next section.

7.5 Translation Algorithms

In a BCI, translation algorithms are those that translate the features extracted from task-specific recordings of EEG into classes of control signals. One control signal associated with a class (one unique signal is associated with each class) works as an input to a computer for execution of a particular action (such as cursor movement). Basic signal analysis steps in BCI have been elaborated in Figure 7.6. We have already discussed about preprocessing (in Chapter 2 and some BCI-specific preprocessing in the beginning of this chapter). We have discussed about BCI-specific feature extraction in the preceding three sections. Now, in this last section of this chapter we will elaborate on the classification of the extracted EEG features in the feature space. Pattern classification or just classification is a vast subject in machine learning (see Duda et al., 2001). Here we will elucidate on only five types, particularly useful for EEG signals.

We will begin with linear classifiers for their simplicity and efficacy.

7.5.1 Fisher's Linear Discriminant

Consider Figure 7.7, in which two different kinds of data points are shown. A projection of those two dimensional points along one fundamental axis will have a clear separation, while a projection along the other will have a mixture that is more difficult to separate out. Exactly this geometric intuition is utilized in Fisher's linear discriminant or FLD. FLD projects a d-dimensional ($d \geq 2$) data set, consisting of two different kinds of data points, on a single line in that d-dimensional space, such that the projected one-dimensional data on that line becomes optimally separated. In many data sets the projected data may still remain quite mixed up and for those data sets FLD is not an effective separator or discriminant.

Suppose that we have a set of $n \cdot d$-dimensional data points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n$ (vectors are written in bold here), n_1 of them are in the subset D_1 labeled ω_1

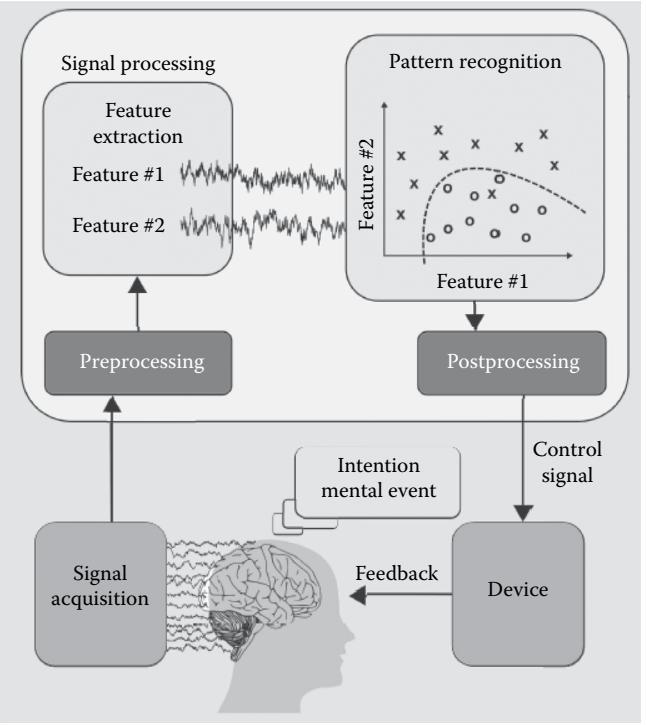


FIGURE 7.6
Block diagram of basic signal analysis steps in the BCI. Brain activity is translated into a control signal for an external device using a sequence of processing stages. The user receives feedback from the device, thereby closing the loop. (Taken from Rao, R.P.N. and Scherer, R., *IEEE Signal Proc. Mag.*, 148, 152, 2010.)

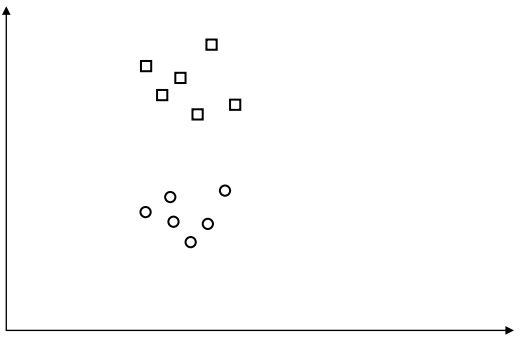


FIGURE 7.7
Two types of data points are shown as squares and circles. If they are projected into the *Y*-axis, they will be clearly separated out from one another, but a projection on the *X*-axis will have a mixture of both in a way that the separation will be more difficult.

and n_2 of them are in the subset D_2 labeled ω_2. Any point $\mathbf{x}_i$ can be projected by the following dot (inner) product

$$y_i = \mathbf{w}^T \mathbf{x}_i, \tag{7.16}$$

where $\mathbf{w}$ is a unit vector signifying only the direction along which $\mathbf{x}_i$ is to be projected. $\mathbf{w}$ is a variable with $\|\mathbf{w}\| = 1$. Our goal will be to arrive at that value of $\mathbf{w}$, which will optimally separate D_1 from D_2 after the mixture of all n $(=n_1+n_2)$ points are projected according to (7.16). That is, we will get points $y_1, y_2, y_3, \ldots, y_n$ separated out into subsets Y_1 and Y_2, where points belonging to D_1 will all be projected onto Y_1 and points belonging to D_2 will all be projected onto Y_2 (like projection of the points in Figure 7.7 onto the Y-axis).

We now turn to the matter of finding the optimum direction $\mathbf{w}$, which will hopefully yield a clear separation of two different data points out of their mixture. One measure of separation between the projected points is the difference of the sample means of Y_1 and Y_2. Let the sample mean for the d-dimensional data points of type i be

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}, \tag{7.17}$$

where $i \in \{1,2\}$. Similarly, let the sample mean for the projected points of the same type be

$$\begin{aligned} \bar{m}_i &= \frac{1}{n_i} \sum_{y \in Y_i} y \\ &= \frac{1}{n_i} \sum_{x \in D_i} \mathbf{w}^T \mathbf{x} = \mathbf{w}^T \mathbf{m}_i. \end{aligned} \tag{7.18}$$

Clearly, $\bar{m}_i$ is the projection of $\mathbf{m}_i$.

It follows that the distance between the projected means is

$$\left| \bar{m}_1 - \bar{m}_2 \right| = \left| \mathbf{w}^T \left(\mathbf{m}_1 - \mathbf{m}_2 \right) \right|. \tag{7.19}$$

To obtain good separation of the projected data it is desirable that (1) the difference between the means $\left| \bar{m}_1 - \bar{m}_2 \right|$ is large, and (2) the standard deviation of Y_1 and that of Y_2 both become small. It is a very elementary yet very profound idea that for a good separation between two classes the mean between them must be large and the variance (or standard deviation) of each of them must be small, which is widely used in statistics. We bring a slightly more generalized notion than variance and call it *scatter* (Duda et al., 2001) defined as

$$\bar{s}_i^2 = \sum_{y \in Y_i} \left(y - \bar{m}_i \right)^2, \tag{7.20}$$

$i \in \{1,2\}$, for the class labeled ω_i. Let us call $\bar{s}_1^2 + \bar{s}_2^2$ the total *within-class scatter* (Duda et al., 2001) of the projected data. FLD employs that linear function $\mathbf{w}^T\mathbf{x}$ for which the ratio

$$J(\mathbf{w}) = \frac{\left|\bar{m}_1 - \bar{m}_2\right|^2}{\bar{s}_1^2 + \bar{s}_2^2} \tag{7.21}$$

is maximum. In this sense $J(\mathbf{w})$ is our *objective function*, which needs to be optimized with respect to the unit vector $\mathbf{w}$ signifying the direction of the projection of the original d-dimensional data (as a mixture of two different types of data).

Let us define $\mathbf{S}_i$ and $\mathbf{S}_w$ as

$$\mathbf{S}_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T, \tag{7.22}$$

$i \in \{1,2\}$ and

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2, \tag{7.23}$$

respectively. Then we can write

$$\begin{aligned} \bar{s}_i^2 &= \sum_{\mathbf{x} \in D_i} \left(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i\right)^2 \\ &= \sum_{\mathbf{x} \in D_i} \left(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i\right)\left(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i\right)^T \\ &= \sum_{\mathbf{x} \in D_i} \mathbf{w}^T (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \mathbf{w} \\ &= \mathbf{w}^T \mathbf{S}_i \mathbf{w}. \end{aligned} \tag{7.24}$$

By (7.23), this implies

$$\bar{s}_1^2 + \bar{s}_2^2 = \mathbf{w}^T \mathbf{S}_w \mathbf{w}. \tag{7.25}$$

Similarly, the separation of the projected mean obeys

$$\begin{aligned} \left(\bar{m}_1 - \bar{m}_2\right)^2 &= \left(\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2\right)^2 \\ &= \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} \\ &= \mathbf{w}^T \mathbf{S}_B \mathbf{w} \end{aligned} \tag{7.26}$$

where

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T. \tag{7.27}$$

We call $\mathbf{S}_w$ the *within-class scatter matrix*. It is proportional to the covariance matrix of the given d-dimensional data set (as a mixture of two different types of data). It is usually nonsingular if $n > d$ (Duda et al., 2001). Likewise, $\mathbf{S}_B$ is called the *between-class scatter matrix*. From (7.27) it is clear that $\mathbf{S}_B$ is the result of the inner product of a vector with itself, and therefore it is a scalar. In particular, given any vector $\mathbf{w}$, $\mathbf{S}_B\mathbf{w}$ is in the direction of $\mathbf{w}$.

From (7.21), (7.24), and (7.26) we can write

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}. \tag{7.28}$$

Note that both the numerator and the denominator of (7.28) are scalar quantities. Therefore, $J(\mathbf{w})$ is a scalar-valued function (of a vector variable). From (7.28) we can write

$$\mathbf{w}^T\left(\mathbf{S}_B\mathbf{w} - J(\mathbf{w})\mathbf{S}_w\mathbf{w}\right) = 0. \tag{7.29}$$

At the maximum of $J(\mathbf{w})$ a small perturbation in $\mathbf{w}$ will keep the value of $J(\mathbf{w})$ unchanged in (7.28). This means for an infinitely many values of $\mathbf{w}$ in a small neighborhood of $\mathbf{w}_0$, where $J(\mathbf{w}_0) = \max(J(\mathbf{w}))$, (7.29) will hold. This is possible only when $\mathbf{S}_B\mathbf{w} = J(\mathbf{w})\mathbf{S}_w\mathbf{w}$ in that neighborhood, in particular,

$$\mathbf{S}_B\mathbf{w} = J(\mathbf{w}_0)\mathbf{S}_w\mathbf{w}. \tag{7.30}$$

Since $J(\mathbf{w}_0)$ is a scalar, (7.30) can be written as

$$\mathbf{S}_w^{-1}\mathbf{S}_B\mathbf{w} = J(\mathbf{w}_0)\mathbf{w}. \tag{7.31}$$

Note that

$$\begin{aligned}\mathbf{S}_B\mathbf{w} &= (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} \\ &= (\mathbf{m}_1 - \mathbf{m}_2)\left\{(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w}\right\}.\end{aligned} \tag{7.32}$$

In $(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w}$, one row vector is multiplied with one column vector of equal dimension. In other words, it is an inner product between two vectors and

therefore is a scalar quantity. So from (7.32) it is clear that $\mathbf{S}_B\mathbf{w}$ is in the direction of $\mathbf{m}_1 - \mathbf{m}_2$. By (7.31) and (7.32) we get

$$\mathbf{S}_w^{-1}\left(\mathbf{m}_1 - \mathbf{m}_2\right)\left\{\left(\mathbf{m}_1 - \mathbf{m}_2\right)^T \mathbf{w}\right\} = J\left(\mathbf{w}_0\right)\mathbf{w}. \tag{7.33}$$

If we ignore the scalar factors in (7.33) we can write

$$\mathbf{w} = \mathbf{S}_w^{-1}\left(\mathbf{m}_1 - \mathbf{m}_2\right), \tag{7.34}$$

where we are concerned only about the direction of $\mathbf{w}$, not about the magnitude. $\mathbf{w}$ in (7.34) gives the direction of FLD. In other words, the d-dimensional hyperplane that separates the data-labeled ω_1 from the data-labeled ω_2 has direction $\mathbf{w}$ (more correctly, the normal to the said hyperplane has direction ratios given by the vector $\mathbf{w}$). Once the direction of the separating hyperplane is determined, the location of the plane needs to be determined. This is done by setting a threshold. The hyperplane is of the form

$$\mathbf{w}^T\mathbf{x} = c, \tag{7.35}$$

where c is a constant. Notice that if data classes ω_1 and ω_2 are linearly separable from each other, then for one of the sets D_i, $i \in \{1,2\}$, say D_1, either of the following must hold

$$\mathbf{w}^T\mathbf{x} \le c, \ \forall \mathbf{x} \in D_1 \quad \text{or} \quad \mathbf{w}^T\mathbf{x} \ge c, \ \forall \mathbf{x} \in D_1, \tag{7.36}$$

where $\forall$ stands for 'for all'.

If the data types ω_1 and ω_2 are not linearly separable from each other, then no matter the threshold chosen, there will always be points on the 'wrong' side of the separating hyperplane given by (7.35). Initially, we can think of the hyperplane passing through the midpoint between the two mean values $\mathbf{m}_1$ and $\mathbf{m}_2$, that is, $\frac{1}{2}\left(\mathbf{m}_1 + \mathbf{m}_2\right)$. Accordingly, the threshold value $c = c_0$ to start with is given by

$$\mathbf{w}^T \frac{1}{2}\left(\mathbf{m}_1 + \mathbf{m}_2\right) = c_0. \tag{7.37}$$

Calculate the true-positive and false-positive rates as given in Section 6.5. Shift the threshold to $c_0 + \delta$ or $c_0 - \delta$, where $\delta > 0$ and recalculate the true-positive and false-positive rates. Set the threshold at an 'acceptable' level of true-positive rate and false-positive rate. This is indeed an ad hoc heuristic, but should work well for most of the EEG applications. The optimum threshold (in some sense) setting is taken care of in the linear support vector machine classifier, which we will discuss in Section 7.5.3. Implementation of

FLD will only involve calculating (7.34) and finding the threshold according to (7.35) on the training data, and then using the $\mathbf{w}$ and c thus obtained on the test data. FLD is a very effective classifier for P300-based BCI applications (Hoffmann et al., 2008).

The next question is what will happen if the number of classes is greater than two? For the C-class problem the natural generalization of FLD involves $C-1$ discriminant functions. So, the projection is from d-dimensional space to $C-1$-dimensional space and it is tacitly assumed that $d \geq C$ (Figure 7.8). The generalization for the within-class scatter matrix is obvious:

$$\mathbf{S}_W = \sum_{i=1}^{C} \mathbf{S}_i, \tag{7.38}$$

where, as before,

$$\mathbf{S}_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \tag{7.39}$$

and

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}. \tag{7.40}$$

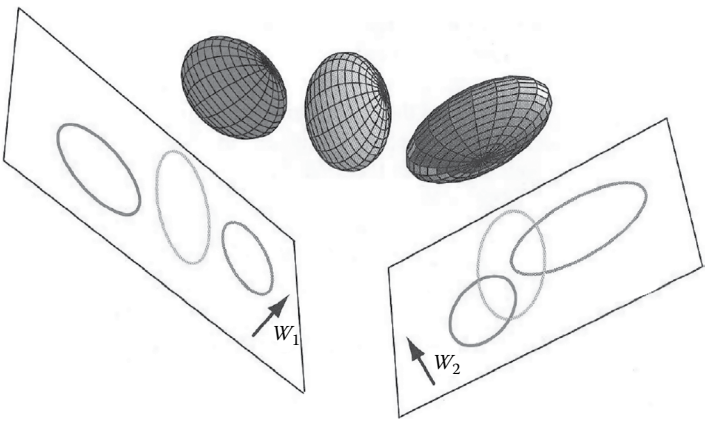


FIGURE 7.8
Three-dimensional data points (one type each in an ellipsoidal cluster) are being projected on two-dimensional planes for maximum separability. Clearly, projection onto the plane on the left (with the direction ratio of the normal being $\mathbf{W}_1$) is more separable than the projection on the plane on the right (with the direction ratio of the normal as $\mathbf{W}_2$). (Taken from Duda, R.O. et al., *Pattern Classification*, 2nd ed., John Wiley & Sons, New York, 2001.)

The proper generalization for $\mathbf{S}_B$ is not quite so obvious (Duda et al., 2001). Suppose we define a *total mean vector* $\mathbf{m}_T$ and the total scatter matrix by $\mathbf{S}_T$ by

$$\mathbf{m}_T = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{C} n_i \mathbf{m}_i \tag{7.41}$$

and

$$\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x} - \mathbf{m}_T)(\mathbf{x} - \mathbf{m}_T)^T. \tag{7.42}$$

Then it follows that

$$\begin{aligned}
\mathbf{S}_T &= \sum_{i=1}^{C}\sum_{\mathbf{x}\in D_i} (\mathbf{x} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m}_T)(\mathbf{x} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m}_T)^T \\
&= \sum_{i=1}^{C}\sum_{\mathbf{x}\in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T + \sum_{i=1}^{C}\sum_{\mathbf{x}\in D_i} (\mathbf{m}_i - \mathbf{m}_T)(\mathbf{m}_i - \mathbf{m}_T)^T \\
&= \mathbf{S}_W + \sum_{i=1}^{C} n_i (\mathbf{m}_i - \mathbf{m}_T)(\mathbf{m}_i - \mathbf{m}_T)^T.
\end{aligned} \tag{7.43}$$

It is natural to define this second term as a general metric between-class scatter matrix, so that the total scatter is the sum of the within-class scatter and the between-class scatter:

$$\mathbf{S}_B = \sum_{i=1}^{C} n_i (\mathbf{m}_i - \mathbf{m}_T)(\mathbf{m}_i - \mathbf{m}_T)^T \tag{7.44}$$

and

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B. \tag{7.45}$$

The projection from a d-dimensional space onto a $C-1$-dimensional space is accomplished by $C-1$ discriminant functions:

$$y_i = \mathbf{w}_i^T \mathbf{x}, \quad i \in \{1, \ldots, C-1\} \tag{7.46}$$

If the y_i's are viewed as a component of a vector $\mathbf{y}$ and the weight vectors $\mathbf{w}_i$'s are viewed as columns of a $d \times (C-1)$ matrix $\mathbf{W}$, then the projection can be written in the matrix form as

$$\mathbf{y} = \mathbf{W}^T \mathbf{x}. \tag{7.47}$$

The samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ projected onto a corresponding to a corresponding set of samples $\mathbf{y}_1, \ldots, \mathbf{y}_n$, which can be described by their own mean vectors and scatter matrices. Thus, if we define

$$\tilde{\mathbf{m}}_i = \frac{1}{n_i} \sum_{\mathbf{y} \in Y_i} \mathbf{y}, \tag{7.48}$$

$$\tilde{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^{C} n_i \tilde{\mathbf{m}}_i, \tag{7.49}$$

$$\tilde{\mathbf{S}}_W = \sum_{i=1}^{C} \sum_{y \in Y_i} \left(\mathbf{y} - \tilde{\mathbf{m}}_i\right)\left(\mathbf{y} - \tilde{\mathbf{m}}_i\right)^T, \tag{7.50}$$

and

$$\tilde{\mathbf{S}}_B = \sum_{i=1}^{C} n_i \left(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}\right)\left(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}\right)^T. \tag{7.51}$$

It is a straightforward matter to show that

$$\tilde{\mathbf{S}}_W = \mathbf{W}^T \mathbf{S}_W \mathbf{W} \tag{7.52}$$

and

$$\tilde{\mathbf{S}}_B = \mathbf{W}^T \mathbf{S}_B \mathbf{W}. \tag{7.53}$$

These equations show how the within-class and between-class scatter matrices are transformed by the projection onto the lower-dimensional space. What we seek is a transformation matrix $\mathbf{W}$ that in some sense maximizes the ratio of the between-class scatter to the within-class scatter. A simple scalar measure of scatter is the determinant of the scatter matrix. If $\mathbf{A}$ is a square matrix and $\mathbf{D}$ is the diagonal matrix obtained due to diagonalization of $\mathbf{A}$, then $|\mathbf{A}| = |\mathbf{B}|$, where $|\mathbf{A}|$ is the determinant of $\mathbf{A}$. So, the determinant of a matrix is the product of the eigenvalues of that matrix. Thus $\left|\tilde{\mathbf{S}}_W\right|$ and $\left|\tilde{\mathbf{S}}_B\right|$ give an estimate of the amount of scatter or variance of the within-class data and between-class data, respectively. Say, for example, if the data is organized in an elliptical form in two dimensions, then the area of the ellipse is πab, where a and b are the semimajor and semiminor axes of the ellipse, respectively. ab is a measure of how the ellipse (of the two-dimensional data points) is stretched or scattered along those two axes. Let us define as objective function on the transformation matrix $J(\mathbf{W})$ as

$$J(\mathbf{W}) = \frac{\left|\tilde{\mathbf{S}}_B\right|}{\left|\tilde{\mathbf{S}}_W\right|} = \frac{\left|\mathbf{W}^T\mathbf{S}_B\mathbf{W}\right|}{\left|\mathbf{W}^T\mathbf{S}_W\mathbf{W}\right|}. \tag{7.54}$$

The problem of finding a rectangular matrix $\mathbf{W}$ that maximizes $J(\mathbf{W})$ is tricky, though fortunately it turns out that the solution is relatively simple (Duda et al., 2001). The columns of an optimal $\mathbf{W}$ are generalized eigenvectors that correspond to the largest eigenvalues in

$$\mathbf{S}_B\mathbf{w}_i = \lambda_i\mathbf{S}_W\mathbf{w}_i. \tag{7.55}$$

A few observations about this solution are in order. First, if $\mathbf{S}_W$ is nonsingular, this can be converted to a conventional eigenvalue problem. However, this is undesirable, since it requires expensive computation of the inverse of $\mathbf{S}_W$. Instead, one can find the eigenvalues as the roots of the characteristic polynomial

$$\left|\mathbf{S}_B - \lambda_i\mathbf{S}_W\right| = 0 \tag{7.56}$$

and then solve

$$\left(\mathbf{S}_B - \lambda_i\mathbf{S}_W\right)\mathbf{w}_i = 0 \tag{7.57}$$

directly for the eigenvectors $\mathbf{w}_i$. Because $\mathbf{S}_B$ is the sum of C matrices of rank 1 or less, and because only $C-1$ of these are independent, $\mathbf{S}_B$ is of rank $C-1$ or less. Thus, no more than $C-1$ eigenvalues are nonzero, and the desired weight vectors correspond to these nonzero eigenvalues (Duda et al., 2001). In general, the solution for $\mathbf{W}$ is not unique. The allowable transformations include rotating and scaling the axes in various ways. These are all linear transformations from a $C-1$-dimensional space onto itself. They leave the objective function $J(\mathbf{W})$ invariant and the classifier unchanged.

One important point is to take $d \geq C$. Only then a d-dimensional data (as a mixture of different types of data) can be projected onto a $C-1$-dimensional hyperplane in the hope of making it nicely separable (i.e., different types of data in the mixture will be separable from each other), as shown in Figure 7.8. Once the data is projected on a low-dimensional hyperplane, some suitable clustering algorithm is needed to classify it into different data types, the mixture of whose original d-dimensional form made up the initial data set $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We will briefly discuss about one such clustering algorithm in Section 7.5.5. Note that after the projection of the data onto an appropriate lower-dimensional hyperplane, as shown in Figure 7.8, classification by clustering may become easier than in the original data set both because of (possible) better separation and dimensionality reduction.

7.5.2 Logistic Regression

Suppose we are designing a linear binary classifier. One popular model of the conditional probability is

$$P\left(Type = 1 \middle| Data = \mathbf{x}\right) = \frac{\exp\left(\beta_0 + \beta^T \mathbf{x}\right)}{1 + \exp\left(\beta_0 + \beta^T \mathbf{x}\right)}, \tag{7.58}$$

where $P(Type = 1 \mid Data = \mathbf{x})$ means the probability distribution of belonging to the Type 1 data cluster, provided the data (point) is $\mathbf{x}$, β_0 is a scalar parameter and β is a vector parameter. Naturally, the conditional probability of belonging to the Type 2 data cluster, provided the data is $\mathbf{x}$, is

$$P\left(Type = 2 \middle| Data = \mathbf{x}\right) = 1 - P\left(Type = 1 \middle| Data = \mathbf{x}\right) = \frac{1}{1 + \exp\left(\beta_0 + \beta^T \mathbf{x}\right)}. \tag{7.59}$$

For a boundary point $\mathbf{x}$ between the two sets of Type 1 and Type 2 data $P(Type = 1 \mid Data = \mathbf{x}) = P(Type = 2 \mid Data = \mathbf{x})$ must hold. We can write it in the *logit* transformation $\log\left(\frac{P}{1-P}\right) = \log 1 = 0$. That is,

$$\log\left(\frac{\exp\left(\beta_0 + \beta^T \mathbf{x}\right)}{1}\right) = 0 \quad \text{or} \quad \beta_0 + \beta^T \mathbf{x} = 0. \tag{7.60}$$

The hyperplane given by (7.60) is the divider or the classifier of the binary data.

Values of β_0 and β are determined by maximum likelihood method. For convenience, let us introduce the notation $p_k(\mathbf{x}_i; \beta) = P(Type = k \mid Data = \mathbf{x}_i; \beta)$, $i \in \{1, \ldots, n\}$, where β is a vector of parameters. For the binary data set it will be convenient to code the two types (Type = 1 and Type = 2) via binary responses $y_i = 1$ when $\mathbf{x}_i$ is Type = 1 and $y_i = 0$ when $\mathbf{x}_i$ is Type = 2, respectively. Let $p_1(\mathbf{x}_i; \beta) = p(\mathbf{x}_i; \beta)$ and $p_2(\mathbf{x}_i; \beta) = 1 - p(\mathbf{x}_i; \beta)$. The joint probability distribution of the all n data points (as a mixture of Type = 1 and Type = 2 data) between clusters $y_i = 1$ and $y_i = 0$ is $\prod_{i=1}^{n} p\left(\mathbf{x}_i; \beta\right)^{y_i} \left(1 - p\left(\mathbf{x}_i; \beta\right)\right)^{1-y_i}$. The log likelihood with respect to the parameter β is given by

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{n} y_i \log p\left(\mathbf{x}_i, \beta\right) + \left(1 - y_i\right) \log\left(1 - p\left(\mathbf{x}_i; \beta\right)\right) \\ &= \sum_{i=1}^{n} \left\{ y_i \beta^T \mathbf{x}_i - \log\left(1 + e^{\beta^T \mathbf{x}_i}\right) \right\}. \end{aligned} \tag{7.61}$$

If $\ell(\beta)$ is to be maximized, then

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}\left(y_i - p(\mathbf{x}_i;\beta)\right) = 0, \tag{7.62}$$

which are called *score* equations (Hastie et al., 2009, p. 120), where x_{ij} is the jth component of the vector $\mathbf{x}_i$. For a d-dimensional space (7.62) is a collection of d, where $j \in \{1, \ldots, d\}$, which will have to be solved numerically, for example, using the Newton–Raphson method.

Note that the first component x_{i1} of $\mathbf{x}_i$ can always be made equal to 1 by $\mathbf{x}_i / x_{i1}$. This implies that the first score equation is

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p(\mathbf{x}_i;\beta). \tag{7.63}$$

We can club together all the d score equations (7.62) into one as $\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_i\left(y_i - p(\mathbf{x}_i;\beta)\right) = 0$, from which we can derive the double derivative as

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T p(\mathbf{x}_i;\beta)\left(1 - p(\mathbf{x}_i;\beta)\right). \tag{7.64}$$

By Newton–Raphson method

$$\beta = \beta^{old} - \left(\frac{\partial^2 \beta}{\partial \beta \partial \beta^T}\right)^{-1}\left(\frac{\partial \ell(\beta)}{\partial \beta}\right). \tag{7.65}$$

Here, the derivative has been evaluated at β^{old}. This should be iterated until there is no significant improvement in the value of β through further iterations.

Logistic regression is a standard classifier for EEG classification in general (Penny et al., 2000; Parra et al., 2005; Tomioka et al., 2007; Prasad et al., 2014). It is usually more reliable and robust (say, in the presence of noise) than the LDA (Hastie et al., 2009, p. 128). In many cases it is almost as efficient as more sophisticated classifiers, such as the support vector machine (Parra et al., 2005, see also subsection 12.3.3 of Hastie et al., 2009), but with lower programming and computational costs. The support vector machine is described right in the next section.

7.5.3 Support Vector Machine

While separating binary data by a linear classifier, if the data is already linearly separable, generally the separating hyperplane (the classifier) can have infinitely many different direction ratios (i.e., the plane can be oriented in infinitely many different ways while still completely separating one type of data from the other by having each type on one side of the plane only). Among all these orientations of the separating hyperplane, which is the direction that will make the probability of misclassification (i.e., at least one data point falling on the wrong side of the hyperplane) the minimum?

Figure 7.9 gives a pictorial representation of a support vector classifier in the simplest case, where the binary data is linearly separable. In the case of a linearly separable data set there exists a thick margin of separation with a nonzero maximum, uniform thickness. Data points, which fall on the boundary of this thick margin, are called *support vectors* (Figure 7.9). The midline of this margin is called the *support vector classifier* (Figure 7.9). It is also known as the 'optimal hyperplane'. Notice that this is the 'safest' position of the classifier, so that it can separate the data with the least chance (probability) that any of the solid dots will fall below the classifier or any of the rings will fall above the classifier. This is called the principle of 'risk (of misclassification) minimization', which is the fundamental premise of the separation algorithm formally known as the *support vector machine* or SVM.

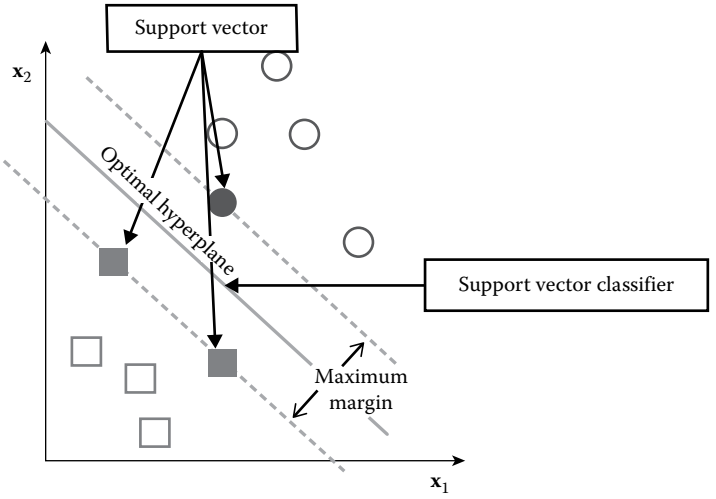


FIGURE 7.9
The simplest linear support vector classifier in a two-dimensional feature space, where the binary data set is already linearly separable. Note that in the case of a linearly separable data set there will be a separation margin with a strictly positive maximum thickness. The support vector classifier or, the optimal (separating) hyperplane will be the midline of this margin and the support vectors will lie on the boundaries of the margin. (Adapted from http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html, accessed April 1, 2017.)

An excellent short description of the SVM with historical notes and some important references is appearing in http://en.wikipedia.org/wiki/Support_vector_machine, which the readers are encouraged to go through. Here we will describe the linear SVM following Hastie et al. (2009).

Case I: Data linearly separable

Consider a separating hyperplane $f(\mathbf{x}) = \beta_0 + \beta^T\mathbf{x} = 0$. The following results regarding the hyperplane are straightforward:

1. For any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ lying in the hyperplane $\beta^T(\mathbf{x}_1 - \mathbf{x}_2) = 0$ is satisfied. β is the vector of direction ratios of the normal to the hyperplane and therefore $\beta^* = \beta/\|\beta\|$ is the unit vector normal to the hyperplane.
2. For any point $\mathbf{x}_0$ in the hyperplane $\beta^T\mathbf{x}_0 = -\beta_0$.
3. Let $\mathbf{x}_0$ be a fixed point in the hyperplane. Then the signed perpendicular distance of any point $\mathbf{x}$ in the space to the hyperplane is

$$\beta^{*T}(\mathbf{x} - \mathbf{x}_0) = \frac{\beta^T}{\|\beta\|}(\mathbf{x} - \mathbf{x}_0) = \frac{\beta^T\mathbf{x} - \beta^T\mathbf{x}_0}{\|\beta\|} = \frac{\beta^T\mathbf{x} + \beta_0}{\|\beta\|} = \frac{f(\mathbf{x})}{\|f'(\mathbf{x})\|}. \quad (7.66)$$

In other words, $f(\mathbf{x})$ is proportional to the signed distance from $\mathbf{x}$ to the hyperplane defined by $f(\mathbf{x}) = 0$.

Let $A = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1,1\}\}_{i=1}^n$, $y_i = 1$ if $\mathbf{x}_i$ belongs to one class and $y_i = -1$ if $\mathbf{x}_i$ belongs to the other. We want to find the optimal separating hyperplane in the sense that it passes through the middle of the widest separating margin between the data sets $A_1 = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^d, y_i = 1\}_{i=1}^{n_1}$ and $A_2 = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^d, y_i = -1\}_{i=n_1+1}^{n}$, while $A_1 \cup A_2 = A$ and $A_1 \cap A_2 = \phi$, ϕ is the empty or the null set.

Now, if a separating hyperplane (not necessarily optimal) is given by $f(\mathbf{x}) = \beta_0 + \beta^T\mathbf{x} = 0$ and if A_1 belongs to the side $\beta_0 + \beta^T\mathbf{x} < 0$ then A_2 will belong to the side $\beta_0 + \beta^T\mathbf{x} > 0$. We can succinctly describe both the cases by $-\sum_{i=1}^n y_i(\beta_0 + \beta^T\mathbf{x}) > 0$. This holds when all data points are correctly classified. In case some points are misclassified we will always want to minimize their number or alternatively, we will want to maximize the number of the rightly classified ones. So we can formulate as, maximize the quantity

$$-\sum_{i\in\Lambda} y_i\left(\beta_0+\beta^T\mathbf{x}\right), \tag{7.67}$$

where Λ is an index set containing indices of the correctly classified data points.

SVM algorithm generalizes (7.67) for the optimally separating hyperplane. Notice that when the data is linearly separable all the points can be correctly classified by a hyperplane. In that case the optimization problem is

$$\begin{aligned}&\max_{\beta,\beta_0,\|\beta\|=1} M\\ &given \quad y_i\left(\mathbf{x}_i^T\beta+\beta_0\right)\geq M, \quad i\in\{1,\ldots,n\}.\end{aligned} \tag{7.68}$$

The set of conditions ensure that all the points are at least a signed distance M away from the decision boundary defined by $\mathbf{x}_i^T\beta+\beta_0=0$. We want to have the largest M and the corresponding parameters β (a vector of parameters) and β_0 (a scalar). Let us write the above inequality (7.68) as

$$\frac{1}{\|\beta\|}\left(\mathbf{x}_i^T\beta+\beta_0\right)\geq M, \tag{7.69}$$

which takes care of the condition $\|\beta\|=1$, so that we do not need to write it separately. Let us rewrite (7.69) as

$$y_i\left(\mathbf{x}_i^T\beta+\beta_0\right)\geq M\|\beta\|. \tag{7.70}$$

Since for any β and β_0 satisfying these inequalities any positively scaled multiple satisfies them too, we can arbitrarily set $\|\beta\|=1/M$ (Figure 7.10). Thus (7.68) is equivalent to

$$\begin{aligned}&\min_{\beta,\beta_0}\frac{1}{2}\|\beta\|^2\\ &given \quad y_i\left(\mathbf{x}_i^T\beta+\beta_0\right)\geq 1, \quad i\in\{1,\ldots,n\}.\end{aligned} \tag{7.71}$$

Note that $y_i\left(\mathbf{x}_i^T\beta+\beta_0\right)\geq 1$ implies $y_i\dfrac{\left(\mathbf{x}_i^T\beta+\beta_0\right)}{\|\beta\|}\geq\dfrac{1}{\|\beta\|}$, where $y_i=1$ or $y_i=-1$. In light of (7.68) $\dfrac{\mathbf{x}_i^T\beta+\beta_0}{\|\beta\|}$ is the distance of $\mathbf{x}_i$ from $\mathbf{x}_i^T\beta+\beta_0=0$, which is greater than or equal to $1/\|\beta\|$ according to $y_i\dfrac{\left(\mathbf{x}_i^T\beta+\beta_0\right)}{\|\beta\|}\geq\dfrac{1}{\|\beta\|}$ or

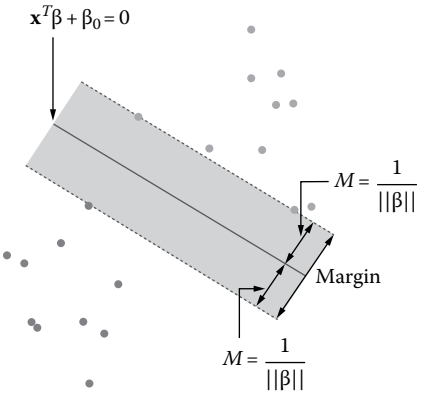


FIGURE 7.10
The support vector classifier $\mathbf{x}^T\beta + \beta_0 = 0$ on a linearly separable binary data set with the maximum separating margin width $2M$, where $M = 1/\|\beta\|$. (Taken from Hastie, T. et al., *The Elements of Statistical Learning Theory: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, 2009.)

$\left\| \frac{\left(\mathbf{x}_i^T\beta + \beta_0\right)}{\|\beta\|} \right\| \geq \frac{1}{\|\beta\|}$. This means the closest data point $\mathbf{x}_i$ to the separating hyperplane $\mathbf{x}_i^T\beta + \beta_0 = 0$ is $1/\|\beta\|$ distance away from the hyperplane. The data point can lie on either side of the hyperplane, that is, the width of the separating margin is $2/\|\beta\|$, whose midline is the separating (optimal) hyperplane. The closest data points to the separating hyperplane are called the support vectors (Figure 7.9).

The solution of (7.71) is known as the *quadratic optimization problem*. The general form of the quadratic optimization problem is optimization of a quadratic form subject to linear constraints exactly as in (7.71). The Lagrange (primal) function to be minimized with respect to β and β_0 is

$$L_P = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i \left(\mathbf{x}_i^T\beta + \beta_0 \right) - 1 \right]. \tag{7.72}$$

Setting the derivatives to zero, we obtain

$$\beta = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \tag{7.73}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{7.74}$$

and substituting these in (7.72) we get the so-called Wolfe dual (Hastie et al., 2009, p. 133)

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k, \quad \text{subject to} \quad \alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{7.75}$$

The solution is obtained by maximizing L_D in the positive orthant, a simpler quadratic optimization problem, for which standard software can be used (Hastie et al., 2009, p. 133). In addition, the solution must satisfy Karush–Kuhn–Tucker conditions, which include (7.73), (7.74), (7.75), and

$$\alpha_i \left[y_i \left(\mathbf{x}_i^T \beta + \beta_0 \right) - 1 \right] = 0, \ \forall i. \tag{7.76}$$

From these we can see that

- If $\alpha_i > 0$, then $y_i \left(\mathbf{x}_i^T \beta + \beta_0 \right) = 1$, that is, $\mathbf{x}_i$ is on the boundary of the thick separating margin
- If $y_i \left(\mathbf{x}_i^T \beta + \beta \right) > 1$, then $\mathbf{x}_i$ is not on the boundary of the separating margin

From (7.73) we see that the solution vector β is defined in terms of a linear combination of the support vectors $\mathbf{x}_i$, which are defined to be on the boundary of the maximally wide separating margin via $\alpha_i > 0$. β_0 is obtained by solving (7.76) for any of the support vectors.

The optimally separating hyperplane produces a function $\hat{f}(\mathbf{x}) = \mathbf{x}^T \hat{\beta} + \hat{\beta}_0$ for classifying new observations ($\hat{g}$ denotes estimation for g):

$$\hat{G}(\mathbf{x}) = sign\ \hat{f}(\mathbf{x}). \tag{7.77}$$

It is intuitively clear that the wider the separating margin is on the test data set, the better will be the classification (in the sense that the chances of misclassification will be lower) in the test data set.

Case II: Data linearly inseparable

So far we have treated only the case where the data set (as a mixture of two types of data) is linearly separable (Figure 7.10). Now we are going to consider the case where the binary mixture is not linearly separable,

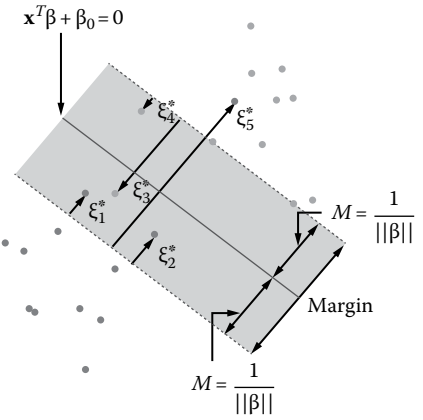


FIGURE 7.11
The binary mixture of data is not linearly separable. The objective function to minimize here is the distance (in some sense) of points falling on the wrong side of a linear classifier from that classifier (the separating hyperplane). (Taken from Hastie, T. et al., *The Elements of Statistical Learning Theory: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, 2009.)

that is, any linear separator will have at least one data point on the wrong side of the separating hyperplane (Figure 7.11). It is also known as the soft margin case.

When the two different types of data sets overlap in the feature space in a linearly nonseparable way, the best linear separator (classifier) will be the one from which the distance of the data points falling on the wrong side is the minimum (Figure 7.11). Let us start with (7.68) and modify the linear constraint as

$$y_i\left(\mathbf{x}_i^T\beta + \beta_0\right) \geq M\left(1 - \xi_i\right), \tag{7.78}$$

$\forall i$, $\xi_i > 0$, $\sum_{i=1}^{n} \xi_i \leq k$, where k is a constant and ξ_i's are called *slack variable*. Each ξ_i is the proportional amount by which the prediction $f\left(\mathbf{x}_i\right) = \mathbf{x}_i^T\beta + \beta_0$ (see Equation 7.77) is on the wrong side of the separating hyperplane $f(\mathbf{x}_i) = 0$. Hence, by $\sum_{i=1}^{n} \xi_i \leq k$, we set a bound for the total proportional amount by which the predictions fall on the wrong side. Note that as long as $\xi_i \leq 1$, $\mathbf{x}_i$ falls on the right side by (7.78). Misclassification occurs when $\xi_i > 1$. So $\sum_{i=1}^{n} \xi_i \leq k$ implies that the total number of training misclassifications cannot exceed k.

As in (7.71) we can drop the norm constraint on β, define $M = 1/\|\beta\|$, and write (7.71) in the equivalent form:

$$\min \frac{1}{2}\|\beta\|^2 \text{ subject to } \begin{cases} y_i\left(\mathbf{x}_i^T\beta + \beta_0\right) \geq 1 - \xi_i, \ \forall i \\ \xi_i \geq 0, \ \sum \xi_i \leq k, \end{cases} \tag{7.79}$$

where k is a constant. This is the usual way in which the support vector classifier is defined for the case where the data is not linearly separable.

By criterion (7.79) we see that the points well inside their class boundary do not play a big role in shaping the boundary. This seems like an attractive property, and one that differentiates it from linear discriminant analysis (Section 7.5.1). In LDA the decision boundary is determined by the covariance of the class distributions and the positions of the class centroids. Logistic regression is more similar to support the vector classifier in this regard (Hastie et al., 2009, p. 419).

Computationally, it is convenient to rewrite (7.79) in the following equivalent form

$$\begin{aligned} &\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\textit{given} \quad \xi_i \geq 0, \ y_i\left(\mathbf{x}_i^T\beta + \beta_0\right) \geq 1 - \xi_i, \ \forall i, \end{aligned} \tag{7.80}$$

where the 'cost' parameter C replaces the constant k in (7.79). The separable case corresponds to $C = 0$ and $\xi_i = 0, \forall i$.

The Lagrange (primal) function is

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left\{y_i\left(\mathbf{x}_i^T\beta + \beta_0\right) - \left(1 - \xi_i\right)\right\} - \sum_{i=1}^{n}\mu_i\xi_i, \tag{7.81}$$

which we minimize w.r.t. β, β_0 and ξ_i. Setting the respective derivatives to zero, we get

$$\beta = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \tag{7.82}$$

$$0 = \sum_{i=1}^{n}\alpha_i y_i, \tag{7.83}$$

$$\alpha_i = C - \mu_i, \ \forall i, \tag{7.84}$$

as all the positivity constraints $\alpha_i, \mu_i, \xi_i \geq 0, \ \ \forall i$. By substituting (7.82) to (7.84) into (7.81) we get the Lagrangian (Wolfe) dual objective function

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, \tag{7.85}$$

which gives the lower bound on the objective function (7.80) for any feasible solution. We maximize L_D subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. In addition to (7.82) to (7.84), the Karush–Kuhn–Tucker conditions include the constraints

$$\alpha_i \left\{ y_i \left(\mathbf{x}_i^T \beta + \beta_0 \right) - 1 - \xi_i \right\} = 0, \tag{7.86}$$

$$\mu_i \xi_i = 0, \tag{7.87}$$

$$y_i \left(\mathbf{x}_i^T \beta + \beta_0 \right) - 1 - \xi_i \geq 0, \tag{7.88}$$

for $i \in \{1, \ldots, n\}$. Together these Equations 7.82 through 7.84 and 7.86 through 7.88 uniquely characterize solution to the primal (7.81) and the dual problem (7.85).

From (7.82) we see that the solution for β has the form

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i, \tag{7.89}$$

with nonzero coefficients $\hat{\alpha}_i$ only for those observations i for which the constraints in (7.88) are satisfied (due to (7.86)). These observations are called support vectors, since $\hat{\beta}$ is represented in terms of them alone. Among these support vectors some will lie on the edge of the thick boundary ($\hat{\xi}_i = 0$ case) and hence from (7.87) and (7.84) will be characterized by $0 < \hat{\alpha}_i < C$. For the rest of the data points, for which $\hat{\xi}_i > 0$ holds, $\hat{\alpha}_i = C$ is satisfied. From (7.86) we can see that any of these support vectors

($\hat{\alpha}_i > 0,\ \hat{\xi}_i = 0$) can be used to obtain $\hat{\beta}_0$. For the sake of numerical stability we actually use the average of all such $\hat{\beta}_0$'s.

Maximizing the dual (7.85) is a simpler convex quadratic programming problem than the primal (7.81) and can be solved using standard techniques (see Gill et al., 1981, p. 177, for example).

Given the solutions $\hat{\beta}$ and $\hat{\beta}_0$, the decision function can be written as

$$\begin{aligned}\hat{G}(\mathbf{x}) &= sign\left(\hat{f}(\mathbf{x})\right) \\ &= sign\left(\mathbf{x}^T\hat{\beta} + \hat{\beta}_0\right).\end{aligned} \tag{7.90}$$

The tuning parameter of this procedure is the cost parameter C.

Here, we have discussed only two simplest forms of SVMs. There is actually a class of SVMs, one each for a particular quadratic objective function to be optimized subject to a set of linear constraints. For more details, see Cristianini and Shawe-Taylor (2000).

Implementation

LIBSVM (Chang and Lin, 2014) developed at the National Taiwan University, Taipei, Taiwan is the popular open-source software implementing various SVMs. See also Chang and Lin (2011). Scikit Learn (http://scikit-learn.org/stable/modules/svm.html) is the Python implementation of LIBSVM. Scikit internally uses LIBSVM to handle all computations (Figure 7.12).

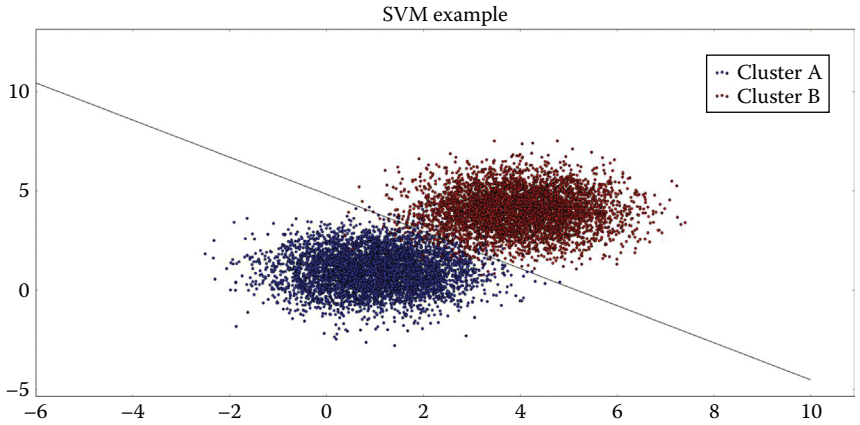


FIGURE 7.12
Implementation in Python of a linear support vector classifier on a two-dimensional binary data set, which is not linearly separable.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import svm

#Creating clusters using a multivariate normal distribution
mean1 = [1,1]
cov = [[1,0],[0,1]]
mean2 = [4,4]
x,y = np.random.multivariate_normal(mean1,cov,5000).T
a,b = np.random.multivariate_normal(mean2,cov,5000).T
data_0 = [list(i) for i in zip(x,y)]
data_1 = [list(i) for i in zip(a,b)]
train = data_0 + data_1  # Training set
label = [0]*5000 + [1]*5000 # Labels
# Training classifier
clf = svm.SVC(kernel='linear') # Linear Kernel, C =1.0
clf.fit(train,label) # fitting training dataset


# finding cordinates of hyperplane for plotting
w = clf.coef_[0]
a = -w[0] / w[1]
xx = np.linspace(-8, 10)
yy = a * xx - (clf.intercept_[0]) / w[1]

#Plotting
data_0 = np.array(data_0)
data_1 = np.array(data_1)
hyperplane=plt.plot(xx,yy,'k-',hold = 'False',label =
  'Hyperplane')
cluster_A=plt.scatter(data_0[:,0],data_0[:,1], hold =
  'True')
cluster_B=plt.scatter(data_1[:,0],data_1[:,1], hold =
  'True',c='red')
plt.legend([hyperplane,cluster_A,cluster_B],['hyperplane',
  'cluster A','cluster B'])

plt.axis('tight')
plt.show()

plt.axis('tight')
plt.show()
```

The P300 speller is based on the concept that in a $N \times N$ matrix of symbols one row and one column have to be highlighted in order to highlight the symbol at their intersection. If the highlighted symbol is the one on which the subject is concentrating then a strong P300 ERP (see Chapter 4) is expected to be detected in the subject's EEG from the occipital lobe. This speller-related P300 is different between the two

cases of when the highlighted symbol is concentrated upon (i.e., it is anticipated) and when the highlighted symbol is not concentrated upon (i.e., it is not anticipated). The amplitude of the EEG signals collected through channels placed in the occipital lobe is the feature. Based on these features, signals have been separated between the events when the anticipated symbol was seen and was not seen, using the SVM represented by (7.80) (Kaper et al., 2004). One additional trick that was used in this work is the kernel method, which we did not discuss in this section. Please see Cristianini and Shawe-Taylor (2000) or Hastie et al. (2009) for the kernel method in SVM.

7.5.4 Neural Network

So far, whatever classifier we have discussed under the Section 7.5 has remained linear. The kernel trick can map a linearly inseparable data mixture (in the feature space) into a linearly separable one in a higher dimensional space than the original feature space. Once the data becomes linearly separable in the higher-dimensional space the data can be separated by a linear SVM represented by (7.71).

A neural network (NN) or an artificial neural network (ANN), on the other hand, is an inherently nonlinear classifier. There are myriad different types of ANNs, but here we will discuss only the back propagation ANN, which has remained by far the most widely used ANN algorithm. A typical back-propagation ANN architecture is shown in Figure 7.13. First, the features are to be extracted from the EEG signals (they can be wavelet coefficients

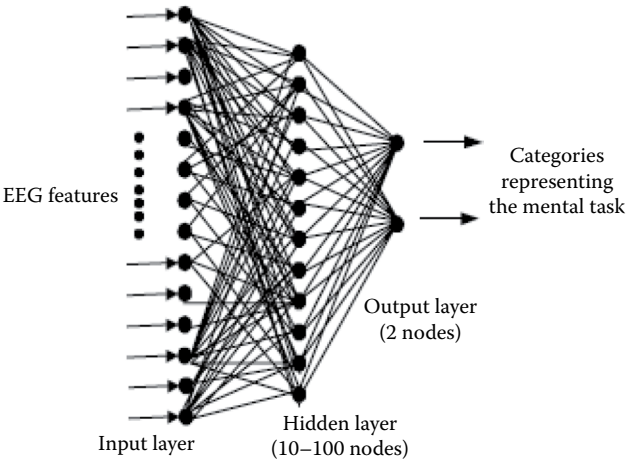


FIGURE 7.13
A typical neural network architecture for EEG-based BCI. (Taken from Huan, N.-J. and Palaniappan, R., *J. Neural Eng.*, 1, 142, 2004.)

or estimated power spectra or any other features). There will be as many input nodes in the ANN as there will be number of features (Figure 7.13). Usually, the number of nodes in the hidden layer (i.e., any layer other than input or output layer, we will soon see that one hidden layer is enough for most purposes), is more than the number of nodes in the input layer. This is because the feature space is mapped into a higher dimensional space where the data can be classified by a linear or a piecewise-linear classifier, which is the fundamental premise of the kernel method. The output layer consists of as many nodes as there are mental tasks into which the EEG signals are to be classified (Figure 7.13).

Let us start with the simplest case. What does a neural network with only one output node (and without any hidden layer) do? After taking the input from the feature space a weighted sum is formed as shown in Figure 7.14. If the feature space is five-dimensional, then an input will be of the form $(x_1, \ldots, x_5)^T$. Let the output node be y_2 and each path joining x_i, $i \in \{1, \ldots, 5\}$ and y_2 signifies a *connection weight* (also known as the *synaptic weight*) w_{i2}. Then $\sum_{i=1}^{5} w_{i2}x_i$ will have a value. Let us denote it by b. The output node y_2 often computes a function like $\exp\left(b - \sum_{i=1}^{5} w_{i2}x_i\right)$, and if it takes a value above a (preset) threshold y_2 yields an output 1 (i.e., the neuron y_2 fires), else it yields an output 0. For suitable values of b and w_{i2} the neuron y_2 can be made to fire if $b > \sum_{i=1}^{5} w_{i2}x_i$, or in other words, if the point represented by $(x_1, \ldots, x_5)^T$ falls on the lower side of the hyperplane $b = \sum_{i=1}^{5} w_{i2}x_i$.

Now consider Figure 7.15. Here, an ANN with one hidden layer is shown. The nodes in the hidden layer approximate the surface that is separating one particular class of objects from the other class(es). Each hidden layer

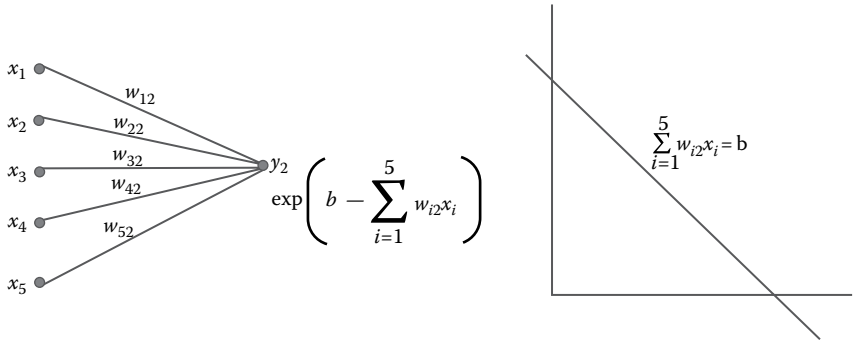


FIGURE 7.14
A single output node ANN without any hidden layer. It generates a hyperplane in the multidimensional feature space, which divides the space into two mutually disjoint halves, into one of which the classified object(s) will fall.

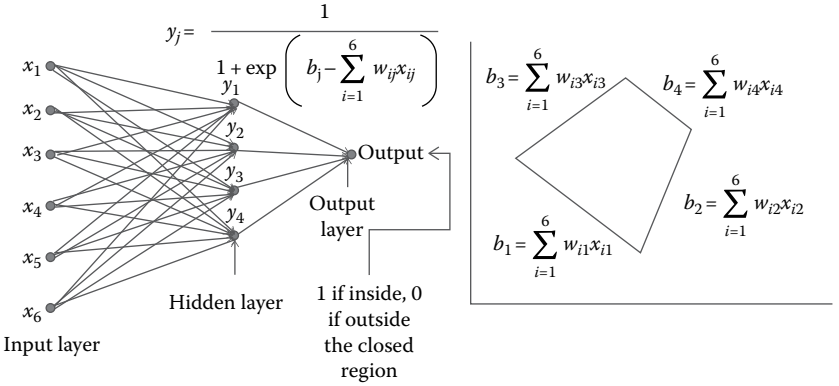


FIGURE 7.15
An ANN with one hidden layer. Each node in the hidden layer determines one hyperplane in the feature space, on one side of which one particular of object is to fall. Ultimately, that class of objects is enclosed by the piecewise hyperplanes as shown on the right side to be separated out from other objects and to be classified into the class they belong to. Likewise the other objects will also be classified.

node of Figure 7.15 calculates as shown in Figure 7.14. The output node in Figure 7.15 computes a conditional statement like (if the points fall on side 1 of plane 1, if the points fall on side 2 of plane 2, if the points fall on side 1 of plane 3, and if the points fall on side 1 of plane 4), then $y_2 = 1$, else $y_2 = 0$. It is intuitively clear that any nonlinear surface separating two different classes of objects one from another can be approximated by this type of piecewise linear approximation approach. But a continuous surface is nothing but a graph of a continuous function in the feature space. In this sense, an ANN with only one hidden layer is able to approximate any continuous function defined on a compact set (a closed and bounded set) in the feature space. This is why an ANN with only one hidden layer is termed as a universal approximator.

An ANN is a supervised learning algorithm. So it needs training on a training set for fixing the parameters b (bias) and w_{i2} (weight). These values are to be fixed on a training set. After being fixed by training, the ANN with these values will have to be applied on the test data set. A training data set is one for which we know which points in the feature space belong to which class. A test data set is one for which this information is unknown, but is expected to be determined by the ANN at hand after its training is over on the training set.

The back propagation algorithm works on the training set in the following way. For each input in the training set the output is found. Then the difference between the actual output (which is known in the training set) and the obtained output is calculated, which is termed as an error. This error is then propagated back to the nodes preceding the output nodes (Figure 7.16) in

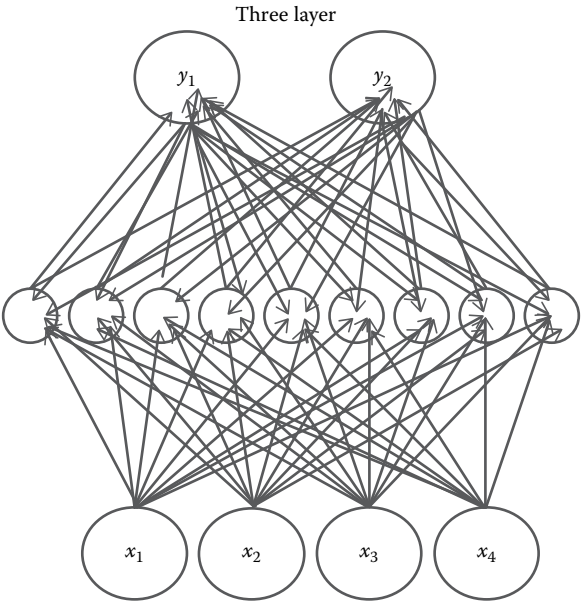


FIGURE 7.16
A back-propagation neural network architecture with one hidden layer. The error back-propagation from the output nodes has been indicated by the backward arrows from the output nodes to the hidden nodes.

the hidden layers, which in turn readjust their calculation by adjusting the values of the connection weights. This adjusting the value of the connection weights in order to better classify the feature vectors (patterns) in the training set is termed as *learning*.

Let us present the learning rule on a per pattern (one single vector in the feature space) basis. Let for a given input pattern $(x_1, \ldots, x_n)^T$ the actual output by the ANN is $(z_1, \ldots, z_m)^T$, whereas the desired output is $(t_1, \ldots, t_m)^T$. The error function to be minimized is

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{m}(t_i - z_i)^2 = \frac{1}{2}\|\mathbf{t} - \mathbf{z}\|^2, \tag{7.91}$$

where

t and **z** are desired and actual output vectors, respectively

w is the weight vector consisting of all the connection weights in the network

The back-propagation learning rule is based on gradient descent. The weights are initialized with random values and then are changed in a direction that will reduce the error

$$\Delta \mathbf{w} = -\eta \frac{\partial \mathbf{J}}{\partial \mathbf{w}}, \tag{7.92}$$

or in the component form

$$\Delta w_{pq} = -\eta \frac{\partial \mathbf{J}}{\partial w_{pq}}, \tag{7.93}$$

where η is the *learning rate* and merely indicates the relative size of the change in the weights (Duda et al., 2001, p. 291). The learning or the change in value of **w** is an iterative process. It is clear from (7.91) that $J(\mathbf{w})$ can never be negative. The negative sign in (7.92) and (7.93) is because $J(\mathbf{w})$ is a decreasing function in each component of **w**. For going from the kth iteration to the $(k + 1)$th iteration in the training set, the following equation

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \Delta \mathbf{w}(k), \tag{7.94}$$

is used in order to update the weight vector from one iteration to the next. The iteration at which $\|\mathbf{t} - \mathbf{z}\| < \in$ is satisfied for some predetermined $\in > 0$ is taken as the last iteration. No further improvement in the error reduction is demanded. Otherwise, it may take too long for the training, which is usually a time-consuming process and has to be performed on a large and diverse data set for better performance. If there are a total of N patterns in the training set, the error function $J(\mathbf{w})$ will take the form

$$J(\mathbf{w}) = \frac{1}{2N} \sum_{i=1}^{N} \left\| \mathbf{t}^i - \mathbf{z}^i \right\|^2, \tag{7.95}$$

where

$\mathbf{t}^i$ is the desired output with the ith input pattern
$\mathbf{z}^i$ is the actual output with the same pattern

In this case, the following has to be achieved with iteration.

$$J(\mathbf{w}) = \sqrt{\frac{1}{2N} \sum_{i=1}^{N} \left\| \mathbf{t}^i - \mathbf{z}^i \right\|^2} < \in, \tag{7.96}$$

where ∈ has been described as above. Calculation of (7.93) for a three-layer network, where $J(\mathbf{w})$ is given by (7.91), is a bit tedious and can be seen in Duda et al. (2001, p. 291). The same can be extended to the case where $J(\mathbf{w})$ is given by (7.95), which will be the case in most of the applications. A meticulously detailed description of a back-propagation neural network and algorithm

can be seen in Hecht-Nilesen (1989). Neural networks can be implemented using the neural network toolbox in MATLAB®.

7.5.5 *k*-Means Clustering

So far we have been discussing supervised learning algorithms, which need to be trained on a training set before being applied on the test set or real-life data. There are machine-learning algorithms outside this paradigm, called unsupervised learning algorithms. The most prominent among the unsupervised learning algorithms are the clustering algorithms. Here we will only discuss about one of the most widely used clustering algorithms, namely, the *k*-means clustering algorithm. k here stands for the number of clusters and will have to be determined by the user right in the beginning of implementing the algorithm. Let n be the total number of data points. Stepwise, the algorithm is described as follows:

1. Arbitrarily determine k points $\mathbf{a}_1^{(0)}, \ldots, \mathbf{a}_k^{(0)}$ in the feature space containing the n data vectors ($n > k$).
2. Let x be an arbitrary data vector, which is closest to $\mathbf{a}_i^{(0)}$ for $i \in \{1, \ldots, k\}$. Then put $\mathbf{x}$ in the ith cluster, that is, in the cluster formed 'centering' $\mathbf{a}_i^{(0)}$ (in some sense). Do it for all data vectors in the feature space. Take the mean of each of the cluster of points. For the ith cluster let the mean be denoted by $\mathbf{a}_i^{(1)}$ for $i \in \{1, \ldots, k\}$.
3. Iterating step 2 $p+1$ iterations, if it is observed that $\mathbf{a}_i^{(p)} = \mathbf{a}_i^{(p-1)}$ for all $i \in \{1, \ldots, k\}$, any further iteration should be stopped at the smallest such p.

Clearly, there will be k clusters centering the vectors $\mathbf{a}_1^{(p-1)}, \ldots, \mathbf{a}_k^{(p-1)}$. k is to be determined by the user and subjective. Also, notice that the clustering is 'hard' in nature in the sense that a given data vector must lie in one cluster. Ambiguity of a data's lying in multiple clusters simultaneously is not allowed, although this may be the case in reality. For example, two different mental states can have largely overlapping, very similar EEG profiles. Such ambiguity can be dealt by allowing 'fuzziness' in cluster design. This exactly gave rise to *fuzzy k-means* (or *c-means*) *clustering*, which we are going to describe below.

It is based on the minimization of the following objective function

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m \left\| \mathbf{x}_i - \mathbf{a}_j^{(0)} \right\|^2, \quad 1 < m < \infty, \tag{7.97}$$

where n, k, and $\mathbf{a}_j^{(0)}$ are as above. u_{ij} is the fuzzy membership of the data vector $\mathbf{x}_i$, $i \in \{1, \ldots, n\}$ in the jth cluster for $j \in \{1, \ldots, k\}$. J_m is updated by updating

membership value u_{ij} and updating the position of the centroid $\mathbf{a}_j^{(0)}$ of the jth cluster for $j \in \{1, \ldots, k\}$. Updating u_{ij} is done as follows:

$$u_{ij} = \frac{1}{\sum_{l=1}^{k} \left(\frac{\left\| \mathbf{x}_i - \mathbf{a}_j^{(0)} \right\|}{\left\| \mathbf{x}_i - \mathbf{a}_l^{(0)} \right\|} \right)^{\frac{2}{m-1}}}, \tag{7.98}$$

whereas $\mathbf{a}_j^{(0)}$ is updated as

$$\mathbf{a}_j^{(0)} = \frac{\sum_{i=1}^{n} u_{ij}^m \mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}^m}. \tag{7.99}$$

This iteration stops according to the following *stopping rule*:

$$\left| u_{ij}^{(r+1)} - u_{ij}^{(r)} \right| < \delta, \tag{7.100}$$

where r signifies number of iteration steps and $\delta > 0$ is the iteration termination criterion. Obviously the smaller the δ, the larger the r will be, and the cost of computation will go up. Fuzzy c-means clustering (FCM) is implemented in MATLAB under the command fcm. Let us summarize the stepwise FCM algorithm below:

1. Initialize $U = [u_{ij}]$ matrix to $U^{(0)}$.
2. At rth iteration: compute the centroids $\mathbf{a}_j^{(r)}$ with $U^{(r)}$

$$\mathbf{a}_j^{(r)} = \frac{\sum_{i=1}^{n} \left\{ u_{ij}^{(r)} \right\}^m \mathbf{x}_i}{\sum_{i=1}^{n} \left\{ u_{ij}^{(r)} \right\}^m}.$$

3. Update $U^{(r)}$ to $U^{(r+1)}$ with

$$u_{ij}^{(r+1)} = \frac{1}{\sum_{l=1}^{k} \left(\frac{\left\| \mathbf{x}_i - \mathbf{a}_j^{(r)} \right\|}{\left\| \mathbf{x}_i - \mathbf{a}_l^{(r)} \right\|} \right)^{\frac{2}{m-1}}}.$$

4. If $\| U^{(p+1)} - U^{(p)} \| < \delta$ for $\delta > 0$ and integral $p > 1$ then stop, else go to Step 2.

The matrix norm in Step 4 may be taken as the Frobenius norm.

One successful application of (a modified form of) FCM was in the automatic identification of different stages of sleep from the scalp EEG collected during sleep (Gath and Geva, 1989). More recently, it has been applied in BCI to classify single-trial EEG associated with finger-movement tasks (Hsu et al., 2012).

References

Bashashati, A., M. Fatourechi, R. K. Ward, and G. E. Birch, A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals, *J. Neural Eng.*, **4**: R32–R57, 2007.

Chang, C.-C. and C.-J. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Int. Syst. Tech.*, **2(3)**: 27, 2011.

Chang, C.-C. and C.-J. Lin, LIBSVM—A library for support vector machines, version 3.20, 2014, available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed April 1, 2017).

Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, U.K., 2000.

Donchin, E., K. M. Spencer, and R. Wijesinghe, The mental prosthesis: Assessing the speed of a P300-based brain-computer interface, *IEEE Trans. Rehabil. Eng.*, **8(2)**: 174–179, 2000.

Duda, R. O., P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, New York, 2001.

Gath, I. and A. B. Geva, Unsupervised optimal fuzzy clustering, *IEEE Trans. Pattern Anal. Mach. Intel.*, **11(7)**: 773–881, 1989.

Gill, P. E., W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, U.K., 1981.

Hallett, M., Movement-related cortical potentials, *Electromyogr. Clin. Neurophysiol.*, **34(1)**: 5–13, 1994.

Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Theory: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, 2009.

Hecht-Nilesen, R., Theory of backpropagation neural network, *Proceedings of IJCNN*, Washington, DC, vol. 1, pp. 593–605, June 18–22, 1989.

Hoffmann, U., J.-M. Vesin, T. Ebrahimi, and K. Diserens, An efficient P300-based brain-computer interface for disabled subjects, *J. Neurosci. Methods*, **167(1)**: 115–125, 2008.

Hsu, W.-Y., Y.-C. Li, C.-Y. Hsu, C.-T. Liu, and H.-W. Chiu, Application of multiscale amplitude modulation features and fuzzy c-means to brain-computer interface, *Clin. EEG Neurosci.*, **43(1)**: 32–38, 2012.

Huan, N.-J. and R. Palaniappan, Neural network classification of autoregressive features from electroencephalogram signals for brain-computer interface design, *J. Neural Eng.*, **1**: 142–150, 2004.

Kaper, M., P. Meinicke, U. Grossekathoefer, T. Linger, and H. Ritter, BCI competition 2003—Data set IIb: Support vector machines for the P300 speller paradigm, *IEEE Trans. Biomed. Eng.*, **51(6)**: 1073–1076, 2004.

Kay, S. M., *Modern Spectral Estimation: Theory & Applications*, Pearson Education Inc., New Delhi, India, 2010.

Makhoul, J., Linear prediction: A tutorial review, *Proc. IEEE*, **63(4)**: 561–580, 1975.

McFarland, D. J., L. M. McCane, S. V. David, and J. R. Wolpaw, Spatial filter selection for EEG based communication, *Electroencepahal. Clin. Neurophysiol.*, **103**: 386–394, 1997.

Nicolas-Alonso, L. F. and J. Gomez-Gil, Brain computer interfaces, a review, *Sensors*, **12**: 1211–1279, 2012.

Parra, L. C., C. D. Spence, A. D. Garson, and P. Sajda, Recipes for the linear analysis of EEG, *NeuroImage*, **28(2)**: 326–341, 2005.

Penny, W. D., S. J. Roberts, E. A. Curran, and M. J. Stokes, EEG-based communication: A pattern recognition approach, *IEEE Trans. Rehabil. Eng.*, **8(2)**: 214–215, 2000.

Pfurtscheller, G. and F. H. Lopes da Silva, Event related EEG/MEG synchronization and desynchronization: Basic principles, *Clin. Neurophysiol.*, **110**: 1842–1857, 1999.

Prasad, P. D., H. N. Halahalli, J. P. John, and K. K. Majumdar, Single trial EEG classification using logistic regression based on ensemble synchronization, *IEEE J. Biomed. Health Inform.*, **18(3)**: 1074–1080, 2014.

Proakis, J. G. and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 4th ed., Pearson Education Inc., New Delhi, India, 2007.

Ramoser, H., J. Muller-Gerking, and G. Pfurtscheller, Optimal spatial filtering of single trial EEG during imagined hand movement, *IEEE Trans. Rehabil. Eng.*, **8(4)**: 441–446, 2000.

Rao, R. P. N. and R. Scherer, Brain-computer interfacing, *IEEE Signal Proc. Mag.*, **148**: 152–155, 2010.

Tomioka, R., K. Aihara, and K.-R. Muller, Logistic regression for single trial EEG classification, *Adv. Neural Inf. Proc. Syst.*, **19**: 1377–1384, 2007.

Wang, Y., S. Gao, and X. Gao, Common spatial pattern method in channel selection in motor imagery based brain-computer interface, *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Shanghai, China, pp. 5392–5395, September 1–4, 2005.

Wolpaw, J. R., N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, Brain-computer interfaces for communication and control, *Clin. Neurophysiol.*, **113**: 767–791, 2002.

Wolpaw, J. R., D. J. McFarland, and T. M. Vaughan, Brain-computer interface research at the Wardsworth Center, *IEEE Trans. Rehabil. Eng.*, **8(2)**: 222–226, 2000.

8

An Overview of fMRI

EEG-fMRI dual recording and analysis is a dominant trend of study these days, which we will focus on in the next and last chapter. In order to make this book largely self-contained, we need to present an overview of fMRI. This we will attempt to accomplish in this chapter. For a more elaborate exposure of functional magnetic resonance imaging (fMRI) starting from the physics of imaging to the computation of image processing, the readers may wish to consult a comprehensive text on it, such as Buxton (2009).

8.1 Magnetic Resonance Imaging

The most abundant molecules in an animal body are the hydrogen molecules, because water constitutes the most major part of the body. For humans of any age and gender, at least 50% of the body is water. About 85% of the human brain is water. All tissues contain water at varied densities. Even human bone contains 10%–15% water. The hydrogen ion in the water molecule works as a single proton spinning about its own axis. Being a charged particle a spinning proton works as a tiny magnet, whose polarity is determined by the direction of spin. In our body, at any point of time, these tiny magnets remain oriented in arbitrary directions, but can be aligned in a specific direction by applying a strong magnetic field (Figure 8.1). Figure 8.1 shows how under an applied magnetic field the otherwise arbitrarily oriented tiny spinning proton magnets get aligned along the direction of the applied field.

The applied strong magnetic field as shown by the thick vertical arrow in the bottom of Figure 8.1 is usually denoted by $\mathbf{B}_0$ in the MRI literature and is measured in Tesla or T, which is a measure of magnetic flux density. 1 T = 10,000 gauss, whereas the Earth's magnetic field varies from 0.25 to 0.65 gauss. fMRI machines of magnetic field strength ranging from 1.5 to 7 T are commonly used for clinical and research purposes. The higher the field strength, the better is the image resolution. With higher field strength some artifacts may also get enhanced.

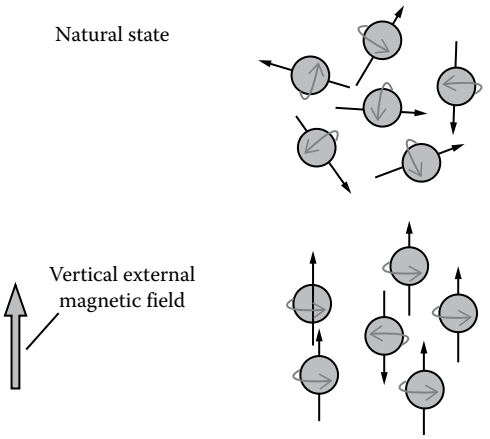


FIGURE 8.1
On top is the natural state of orientation of the single proton magnets. The direction of arrow along the axis of the proton indicates the north pole, and the curve along the surface of the sphere indicates the direction of the spin. In the bottom, under the vertical external magnetic field, the spinning proton magnets have been aligned along the direction of the applied field. (Taken from Kandel, E.R. et al. (eds.), *Principles of Neural Science*, 5th ed., McGraw Hill, New York, 2013.)

8.1.1 T_1-Weighted Imaging

Next, a radio frequency (RF) pulse is applied perpendicular to $\mathbf{B}_0$ (Figure 8.2) for a very short duration (a few milliseconds). Due to the pulse, the tiny proton magnets get tilted from their aligned position parallel to $\mathbf{B}_0$, as shown in Figure 8.2. As soon as the RF pulse is stopped, the proton magnets start to come back to their original positions. The time taken to come back is called T_1. At a fixed instance of time not exceeding T_1, in different tissues different proportion of protons come back to their original positions.

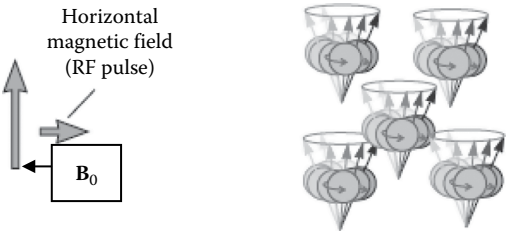


FIGURE 8.2
A horizontal magnetic field perpendicular to $\mathbf{B}_0$ is applied transiently, like a pulse, in order to tilt the spinning protons away from their alignment parallel to $\mathbf{B}_0$. They then keep rotating round their original vertical alignment direction at a fixed frequency, which is dependent on the strength of $\mathbf{B}_0$. This rotation is called precession. (Taken from Kandel, E.R. et al. (eds.), *Principles of Neural Science*, 5th ed., McGraw Hill, New York, 2013.)

Due to the RF pulse, the axis of the proton, around which it is rotating, gets tilted as shown in Figure 8.2. The proton wobbles around the $\mathbf{B}_0$ magnetic field like a spinning top (Figure 8.2), which is called *precession*. Precession creates a rotating magnetic field that changes in time, which, according to Faraday's law, generates an electric current. It is this electric current that is measured in magnetic resonance imaging (MRI). So, if an electrical image of the tilted protons is taken across the tissues at a particular instance within the T_1 duration, an image with different contrasts for different tissues results (Figure 8.3), and this image is called T_1-weighted MR image.

Now let us elaborate the techniques a little further. Our body is full of spinning protons. When we place these tiny magnets in a strong magnetic field (called $\mathbf{B}_0$), some will tend to align in the direction of the magnetic field and some will tend to align in a direction opposite to the magnetic field. The magnetic fields from many protons will cancel out, but a slight excess of the protons will be aligned with the main magnetic field, producing a *net magnetization* that is aligned parallel to the main magnetic field (it is produced by ~0.01% excess protons whose axis of precession is aligned along $\mathbf{B}_0$, and the remaining ~99.99% are aligned equally along and opposite to $\mathbf{B}_0$ and therefore

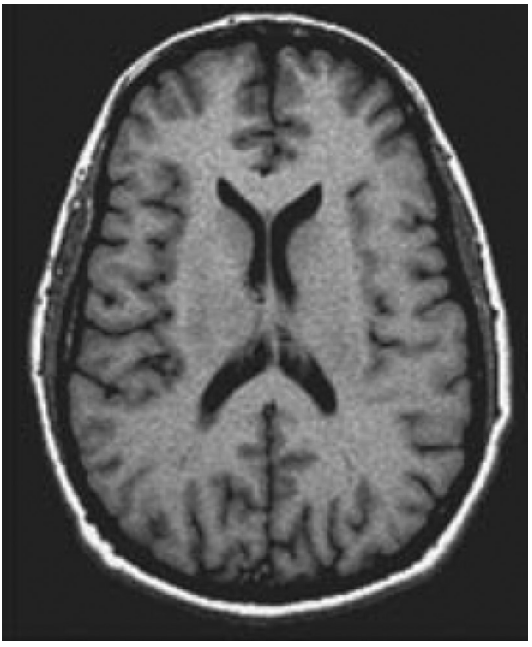


FIGURE 8.3
A T_1-weighted image of a two-dimensional brain slice. Images are taken of cross sections of the brain, a few millimeters apart. Finally, all these two-dimensional images are put together in the order they were taken in order to build up the three-dimensional image of the whole brain. (Taken from Buxton, R.B., *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.)

get cancelled by mutual interactions). This net magnetization becomes the source of our MR signal and is used to produce MR images (Pooley, 2005).

The gyromagnetic ratio has been explained in Figure 8.4. It is nucleus dependent, for example, for the hydrogen nucleus or single proton it is 42.6 MHz/T. So in a 1.5 T machine (i.e., the main magnetic field $\mathbf{B}_0$ is 1.5 T strong), the precession frequency of the protons will be 63.9 MHz, which is in the radio frequency (RF) range. If a RF pulse is given at the same frequency (63.9 MHz) to the precessing protons in a direction perpendicular to $\mathbf{B}_0$, as shown in Figure 8.2, then it will be a *resonance*, where the energy of the RF pulse will be utilized most efficiently to perturb the precessing protons. In fact, the pulse will be given for such a duration that the vertical axis of the rotating (precessing) protons becomes completely flat or horizontal, as shown in Figure 8.5. After the RF pulse is stopped the axis of

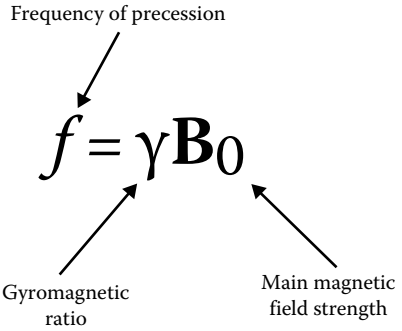


FIGURE 8.4
The frequency of precession or the rate of rotation of the proton magnet around the vertical axis as shown in Figure 8.3 is linearly proportional to $\mathbf{B}_0$. The frequency of precession f in a given field $\mathbf{B}_0$ is called *Larmor frequency*. $f = \gamma \mathbf{B}_0$ is called *Larmor equation*, where γ is called the *gyromagnetic ratio*. (Taken from Pooley, R.A., *RadioGraphics*, 25(4), 1087, 2005.)

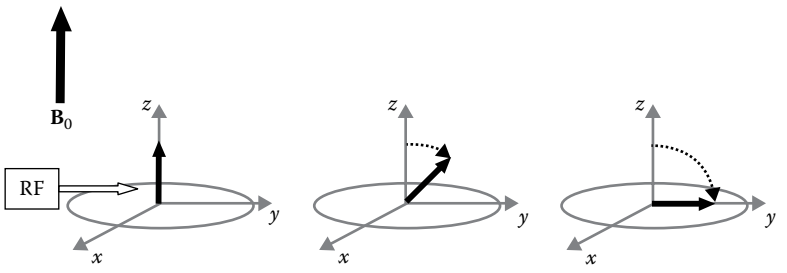


FIGURE 8.5
The main magnetic field $\mathbf{B}_0$, shown by a thick black arrow, is always vertical. The horizontal RF pulse is shown by a white arrow. During the pulse, the precessing proton falls in such a way that its vertical axis of rotation along *z*-axis, shown by a small black arrow, gradually becomes flat or horizontal (right plot). (Taken from Pooley, R.A., *RadioGraphics*, 25(4), 1087, 2005.)

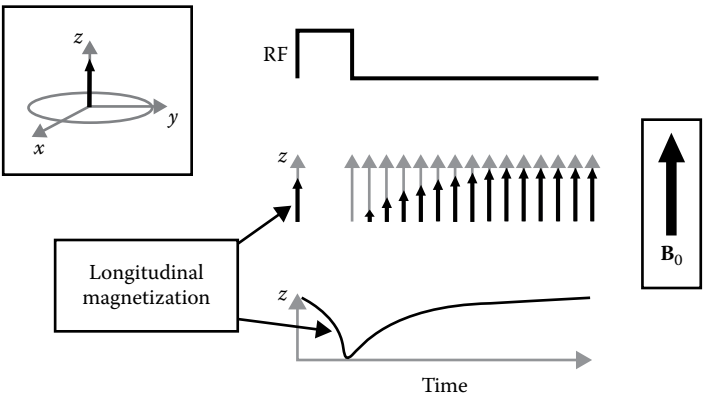


FIGURE 8.6
The black arrow in the top left panel is the axis of rotation of a precessing proton. After the RF pulse is stopped, the axis of rotation of the proton comes back to the vertical position to realign itself to $\mathbf{B}_0$ once again. The time of fall of the axis under the RF pulse has been shown by the left cusp on the bottom plot, and time of recovery back to the vertical position from the flat horizontal position has been shown in the right elongated cusp in the bottom plot. This time plotted in the elongated cusp is the T_1 relaxation time (further elaborated in Figure 8.7). The change in the magnetic intensity of the precessing proton along the vertical direction has been shown in the middle plot. (Adapted from Pooley, R.A., *RadioGraphics*, 25(4), 1087, 2005.)

rotation of the precessing proton starts coming back to the original vertical position to realign itself along $\mathbf{B}_0$ as shown in Figure 8.6. The recovery time taken by the proton for the realignment to vertical position is known as the T_1 relaxation time as shown in Figure 8.6. This has been further elaborated in Figure 8.7.

Different tissues in the body show different resistances to this realignment of the precessing proton axis to its initial vertical position along the $\mathbf{B}_0$. When 63% of the original magnetization of the proton is gained during the realignment process (Figure 8.7), a snapshot or 'photograph' is taken. This photograph or image is called the T_1-weighted image of the tissue (Figure 8.3). This is no visible light photography image, but a magnetic resonance image. We will see shortly how this image is taken converting the net magnetic field of the precessing protons into electric current.

In Figure 8.8, we can see how the contrast of T_1-weighted image varies across different brain tissues. At a moment when the contrast between adjacent tissue layers, such as white matter, gray matter, and cerebrospinal fluid (CSF), is optimum, the T_1-weighted image will give the clearest image of the brain. Note that this is a structural image only. No functional information is captured in this image. The right graph of Figure 8.8 gives an intuitive justification of why 63% recovery of the original magnetization of the precessing proton is chosen for taking the T_1-weighted snapshot.

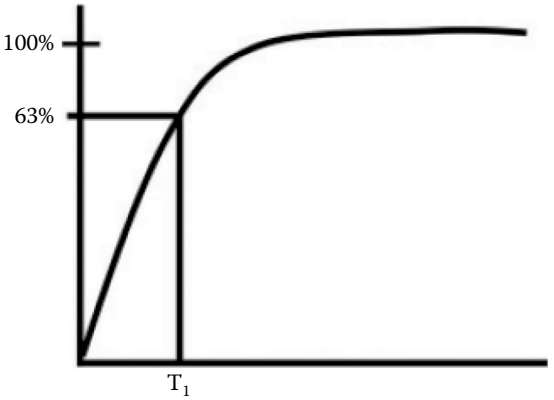


FIGURE 8.7
Elaboration of the T_1 relaxation time as represented by the abscissa of the right cusp in the bottom plot of Figure 8.6. T_1 is a characteristic of tissue and is defined as the time taken for the longitudinal magnetization to grow back to 63% of its final value. (Taken from Pooley, R.A., *RadioGraphics*, 25(4), 1087, 2005.)

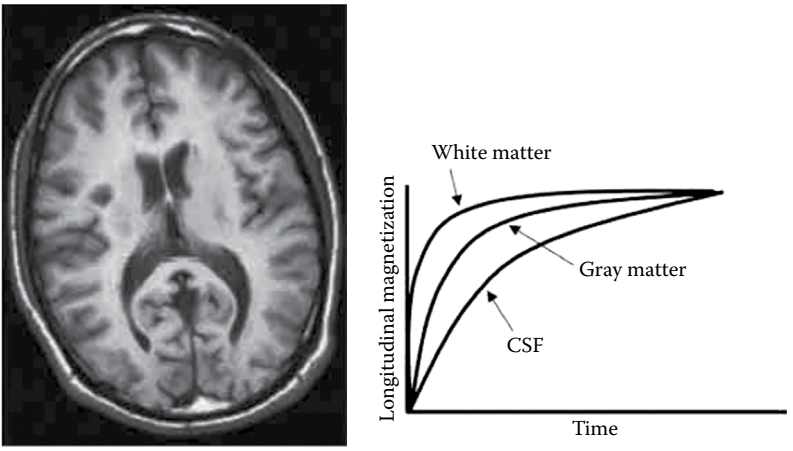


FIGURE 8.8
T_1-weighted contrast. Different tissues have different rates of T_1 relaxation. If an image is obtained at a time when the relaxation curves are widely separated, the T_1-weighted contrast will be maximized. (Taken from Pooley, R.A., *RadioGraphics*, 25(4), 1087, 2005.)

8.1.2 T_2-Weighted Imaging

T_2-weighted imaging can be thought as a 'negative' of T_1-weighted imaging, in the sense that, the tissues appearing bright in T_2 will appear dark in T_1 and vice versa. When the protons are precessing along $\mathbf{B}_0$, they are all phase locked, that is, precessing at a constant phase difference (0 or π depending on the direction of the proton's magnetic field; if it is same as $\mathbf{B}_0$, then the

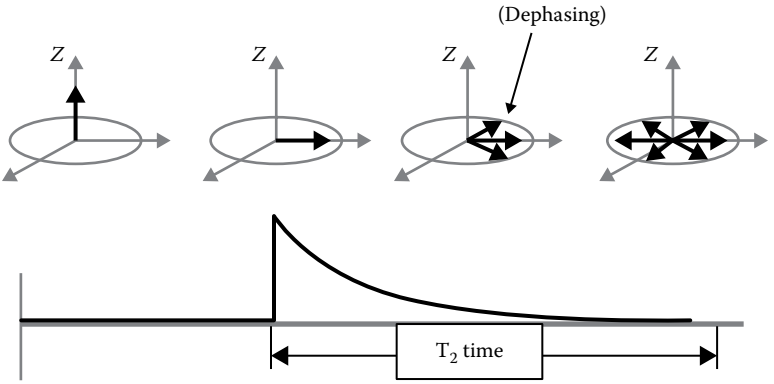


FIGURE 8.9
When the magnetic field due to the precessing protons are parallel to the main magnetic field (vertical), the protons rotate in a phase-locked manner with each other. The net magnetization due to the precessing protons remains highest at that state. After the RF pulse tilts them away from the vertical axis, the protons start becoming dephased among themselves (sequences in the top plot from left to right). Naturally, the net magnetization goes down as shown in the bottom plot. T_2 is the time taken for the net magnetization to become zero from its highest value as shown in the bottom plot. (Taken from Pooley, R.A., *RadioGraphics*, 25(4), 1087, 2005.)

difference is 0, if it is opposite to $\mathbf{B}_0$ then the difference is π) with each other. Immediately after applying the RF pulse perpendicular to $\mathbf{B}_0$ they may still be phase locked, but quickly afterwards they become dephased or phase desynchronized with each other (Figure 8.9). Here we are concerned about the magnetic intensity due to the precessing protons along the *xy*-plane (more generally perpendicular to the *z*-axis), along which the magnetic intensity going down due to dephasing has been shown in the top diagrams. The bottom plot shows how this net magnetization evolves during the T_2 time. The net magnetization goes down from maximum to zero during the T_2 time. Ideally, T_2 time should be infinite, but for all practical purposes it is considered finite. During T_1 time the net magnetization perpendicular to *z*-axis goes up, whereas the net magnetization during the T_2 time goes down. This is the reason behind the T_1-weighted image appearing like a negative of T_2-weighted image and vice-versa. Like the T_1-weighted snapshot protocol there is a T_2-weighted snapshot protocol as well as shown in Figure 8.10. Different tissues allow different times for dying down of the net magnetization due to dephasing and therefore show different contrasts at any given time instant T_2. The convex cusp of Figure 8.10 represents the net magnetization decay along T_2 for one particular tissue only. It is intuitively clear that the approximately 37% intensity point on the cusp will be the point where the tissue contrasts will be maximally separated (Figure 8.11, also compare with Figure 8.8). Therefore, a snapshot at this instant will give the best tissue contrasts and the best quality image (Figure 8.11). In Table 8.1,

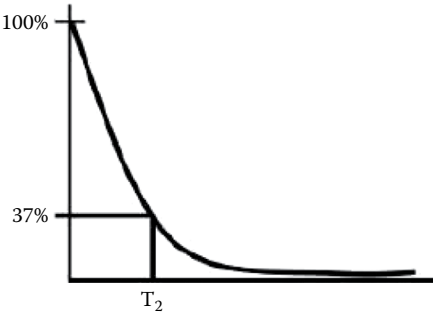


FIGURE 8.10
Here is a magnified version of the T_2 time vs. net magnetization plot as shown in Figure 8.9. This plot is for one particular type of tissue only. Notice that the 37% net magnetization intensity point on the cusp is likely to give the best or near-best tissue contrasts between multiple tissues (Figure 8.11). (Taken from Pooley, R.A., *RadioGraphics*, 25(4), 1087, 2005.)

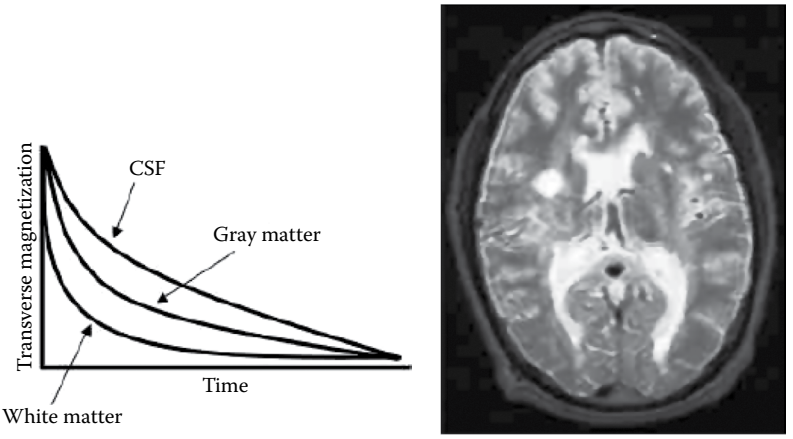


FIGURE 8.11
Just like Figure 8.8, but for the T_2-weighted image. (Taken from Pooley, R.A., *RadioGraphics*, 25(4), 1087, 2005.)

TABLE 8.1

T_1, T_2 Time Constants in Milliseconds at 1 Tesla for Different Tissues in the Brain

Tissues	T_1	T_2
Fat	241	85
White matter	683	90
Gray matter	813	100
CSF	2500	1400

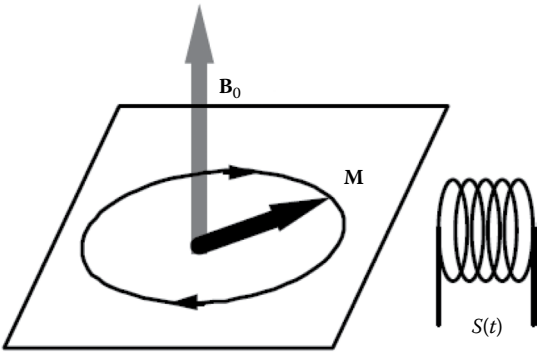


FIGURE 8.12
The net magnetization **M** along *xy*-plane induces a voltage $S(t)$ in the coil placed nearby. $S(t)$ is processed by specialized software to generate a visible image. (Taken from Noll, D.C., A primer on MRI and functional MRI, version 2.1, Department of Biomedical Engineering and Radiology, University of Michigan, Ann Arbor, MI, 2001, available at http://www.cs.ucsb.edu/~mturk/imaging/Misc/MRI%20primer.pdf.)

T_1 and T_2 time constants (i.e., time taken to decay the net magnetization perpendicular to the *z*-axis from maximum intensity to zero intensity) for different brain tissues are shown.

Finally, to generate a visible two-dimensional T_1 or T_2 image like in Figure 8.8 or Figure 8.11, respectively, the net magnetization perpendicular to $\mathbf{B}_0$ is to be 'harvested'. It is captured at the *xy*-plane in different *z*-locations. One location gives a slice like in Figure 8.8 or in Figure 8.11. Placing several such slices one above another in successive order a three-dimensional MR image (MRI) of the whole brain is generated. How then is the net magnetization, say **M** (Figure 8.12) on a *xy*-plane harvested? It is done by simply placing a coil, whose axis is along the plane. Of course, for better signal acquisition there may be multiple such coils surrounding the circle in Figure 8.12. Due to the variation in the magnetic field **M** a voltage $S(t)$ will be induced in the coil (Figure 8.12). Collecting electrical voltage signals $S(t)$ from the coils surrounding the circle in Figure 8.12 and processing them with appropriate software a visible T_1- or T_2-weighted image slice for that particular *z*-location is generated.

8.1.3 Spatial Localization

MR images are of great clinical and research value not only for their detail and clarity, but also for their precise spatial localization of tissue structures. Modern MR scanners can localize tissue shapes with millimeters lower resolution. In order to understand such precise spatial localization we must delve deeper into the MRI acquisition techniques. The whole process may be divided into three subprocesses, namely, *slice selection*, *frequency encoding*, and *phase encoding*.

Slice selection or location along the z-axis.

Conventionally, the z-axis is taken along the main magnetic field $\mathbf{B}_0$, as shown in Figure 8.13. $\mathbf{B}_0$ is aligned along the axis of the huge cylindrical core of the MRI machine. The cylindrical core is the magnet producing $\mathbf{B}_0$. The core is made up of a superconducting material and needs to be kept at a very low temperature by liquid helium. The large hole inside the cylinder is called the *gantry*, inside which the human subject is pushed while lying on a movable stretcher attached to the MR machine (Figure 8.13). In this sense $\mathbf{B}_0$ is aligned along the length of the body.

A uniform magnetic field is not useful for localization. So a gradient or variation is to be applied. This is applied as

$$B_z = \mathbf{B}_0 + G_z z, \tag{8.1}$$

where G_z is piecewise constant (see Figure 8.15 for how it varies), known as the slope or the *gradient* of the strength of the magnetic field along the z-axis. G_z is chosen in a way that $G_z z$ amounts to only a few tens of milliteslas at the maximum. Now from Larmor's equation (Figure 8.4) we know that the frequency $f(z)$ is given by

$$f(z) = \gamma B_z = \gamma(\mathbf{B}_0 + G_z z), \tag{8.2}$$

where γ is the gyromagnetic ratio of single protons. Notice that by virtue of Larmor's equation frequency $f(z)$ has become a function of spatial location z. Now, if the slice normal to the z-axis in Figure 8.13 is thin enough (usually 1–10 mm thick), we can assume the magnetic field strength B_z remains almost

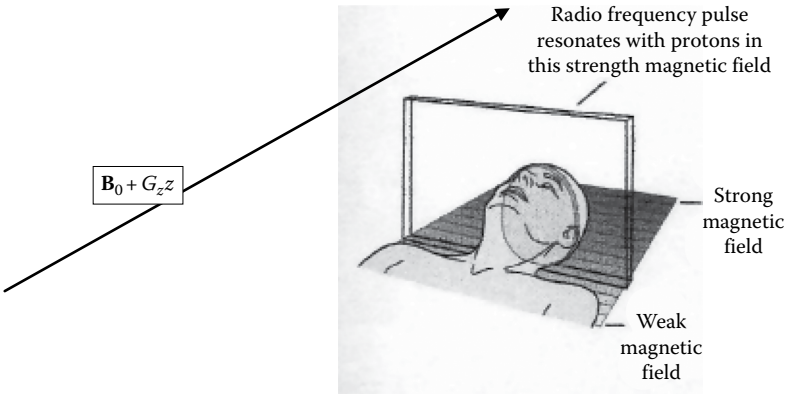


FIGURE 8.13
A schematic view of slice location inside an MRI gantry and the magnetic field gradient along the main magnetic field $\mathbf{B}_0$ (taken along the z-axis) indicated by the arrow. The magnetic field gradient along the z-axis is $\mathbf{B}_0 + G_z z$, where G_z is piecewise constant (see Figure 8.15). (Taken and modified from Kandel, E.R. et al. (eds.), *Principles of Neural Science*, 4th ed., McGraw Hill, New York, 2000.)

constant across the slice. An RF pulse of frequency $f(z)$ (as in Equation 8.2) is applied only horizontal to this slice along the x-axis (this slice is considered as the xy-plane). This RF pulse is then utilized to take a T_1 or T_2 image of this slice. Of course, this process has to be repeated many times over for a better image.

The technical details of imaging the located slice (as in Figure 8.13) can be divided into two paradigms, namely, frequency encoding and phase encoding.

Frequency encoding: This notion is for taking a T_1 or a T_2 image along the x-axis. We have already seen that an RF pulse of frequency $f(z)$ is applied along x-axis on the xy-plane slice in Figure 8.13. Subsequently, a magnetization gradient $B_x = G_x x$ is applied along a single line at a time parallel to the x-axis (Figure 8.14), where G_x is piecewise constant (see Figure 8.15 for how it varies). Without the gradient field the frequency along any line parallel to the x-axis is uniform (upper panel of Figure 8.14). But with a gradient field $B_x = G_x x$ along the line the frequency of precession varies according to the Larmor's equation (lower panel of Figure 8.14).

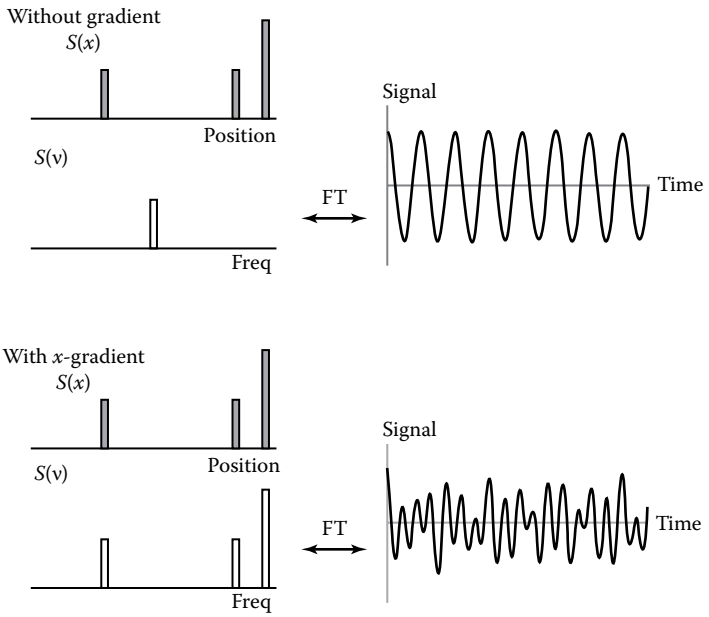


FIGURE 8.14
Schematic diagram of frequency encoding. There are three small signal sources (collection of precessing protons) on a line parallel to the x-axis without any magnetic gradient along the line (top panel) and with a magnetic gradient varying with x (bottom panel). Since the frequency varies linearly with x, a Fourier transform (FT) on the time domain electrical signal, collected by the method shown in Figure 8.12, appearing on the right side of the bottom panel, will give the precise spatial location on the x-axis. (Taken from Buxton, R.B., *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.)

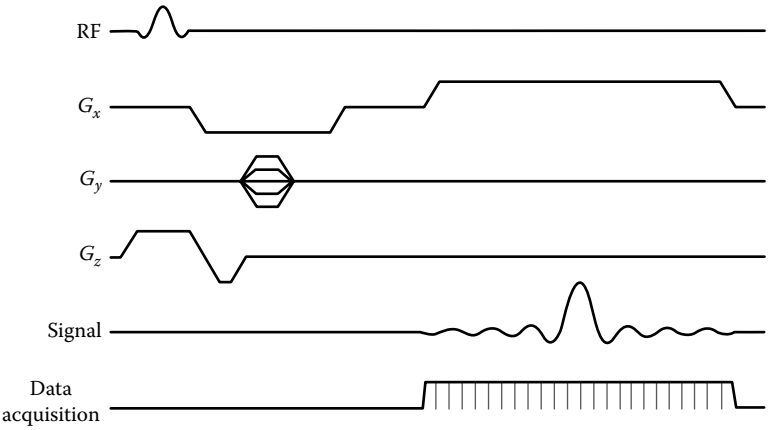


FIGURE 8.15
The order of the pulse sequences along different spatial directions for an MR imaging. During the RF excitation pulse a gradient along *z*-axis is applied (slice selection). During read-out (positive part of the gradient) of the signal a gradient along *x*-axis is applied (frequency encoding). Between these two gradient pulses a gradient pulse along *y*-axis is applied. The amplitude of this last pulse is stepped through a different value each time the pulse is repeated (phase encoding). Typically, 128 or 256 phase-encoding steps (repeats of the pulse along *y*-axis) are required to collect sufficient information to reconstruct an image. Also shown is the time domain signal that results and its digitization. (Taken from Buxton, R.B., *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.)

There are three small signal sources on a line parallel to the x-axis on the selected slice (Figure 8.14). The sources are a collection of precessing protons. How to find their exact location? A magnetic gradient $B_x = G_x x$ (again only a few tens of milliteslas at the most) is to be applied along the line, which will have a frequency representation $f(x) = \gamma B_x = \gamma G_x x$ by the Larmor equation. Note that data acquisition is always during the constant positive value of G_x, as shown in Figure 8.15. It is clear from Figure 8.14 that if we perform a Fourier transformation on the time domain signal in the right bottom corner of the figure we get $f(x)$. Of course, the signal will have to be discretized first so that the discrete Fourier transform (fast Fourier transform or FFT to be more precise) can be applied to it. From $f(x)$ by the Larmor equation we get the position x.

Phase encoding: This is the technique by which the location along the y-axis is selected. Localization in 2D cannot be performed by simply turning on two gradients simultaneously, one along the x-axis and the other along the y-axis. This will lead to a single resultant gradient direction somewhere between the x- and y-axes and will not help in localizing along any of them. So pulses will have to be applied one after another. The order of the pulse sequences along different spatial axes has been shown in Figure 8.15. The RF pulse and the gradient pulse along the z-axis are applied first and simultaneously in order to select the slice as shown in Figure 8.13. Then a gradient pulse along x-axis

is applied (the initial negative gradient is not to be considered for image acquisition) for the frequency encoding as shown in Figure 8.14. In between the pulses along *z*- and *x*-axes applied a pulse along the *y*-axis as shown in Figure 8.15. Note that this pulse is applied for a relatively shorter duration and multiple times (typically 128 or 256 times for a good-quality MR image). Each time the intensity of the pulse is different from that of any other time during a whole two-dimensional slice image acquisition (as appearing in Figure 8.8 for a T_1-weighted image and Figure 8.11 for a T_2-weighted image).

In Figure 8.16, a gradient pulse G is shown in the top panel, whose magnitude × duration on the positive side is equal to the magnitude × duration on the negative side. When protons are precessing in a uniform field

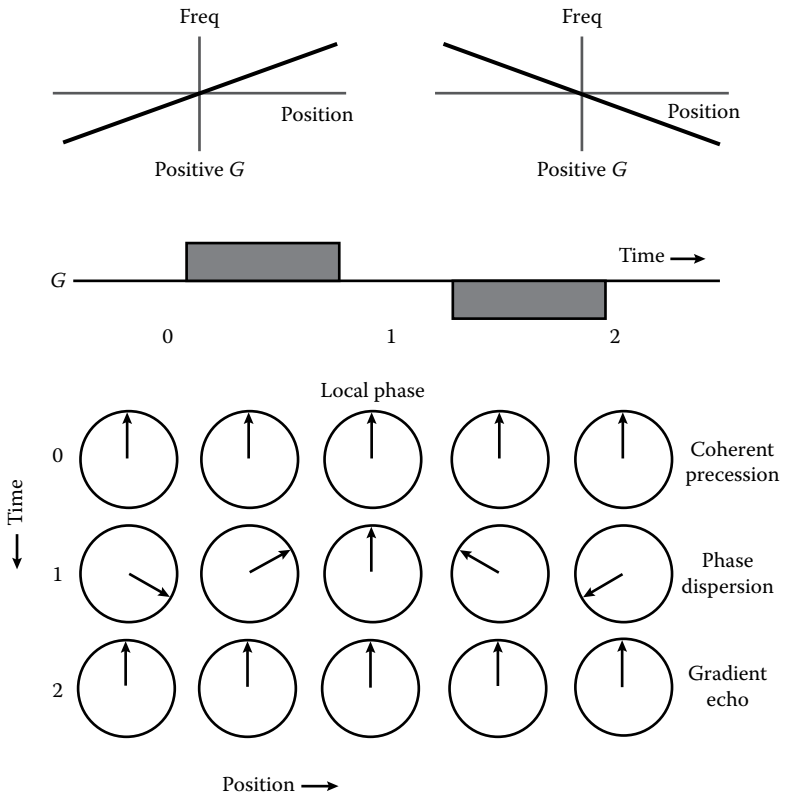


FIGURE 8.16
Effect of gradient pulse. In a uniform field, spins precess at the same rate and remain in phase (time point 0). A gradient field produces a linear variation of the precession rate with position (top). At the end of the first gradient pulse (time point 1), the local phase angle varies linearly across the object (phase dispersion). A gradient pulse of opposite sign and equal area reverses these phase offsets and creates a gradient echo when the spins are back in phase (time point 2). (Taken from Buxton, R.B., *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.)

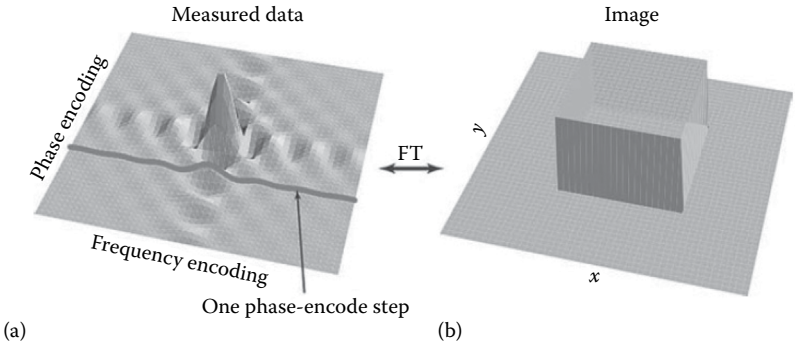


FIGURE 8.17
(a) Slice imaging by frequency and phase encoding. Frequency encoding along x direction is done one line at a time. The position of each line is decided by phase encoding. (b) Fourier-transform two-dimensional signal, which gives the precise location in the xy-plane. (Taken from Buxton, R.B., *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.)

phase differences among them is constant (top row in the bottom panel of Figure 8.16). After applying the gradient pulse with positive amplitude, transverse to the axis of rotation of the proton, the protons become out of phase among themselves (middle row in the bottom panel of Figure 8.16). When the pulse is stopped protons still precess, but their phase differences remain fixed. But when a pulse is applied in the opposite direction with equal magnitude and duration of the first pulse, the precessing protons go back to their original position (bottom row of the bottom panel in Figure 8.16).

The gradient pulse G_y is applied before the gradient pulse G_x (with effect $B_x = G_x x$ on position x) and perturbs the phase of the precessing protons. When the pulse is stopped, the phase as changed by the pulse remains in that state. Then when the gradient along the x-axis is turned on in the forward and reverse directions with equal duration and amplitude in each cycle, as shown in Figure 8.16, during the data acquisition from the signal (Figure 8.15) the x position is localized along a line parallel to the x-axis in the slice plane (frequency encoding). A new gradient pulse G_y (different in magnitude than the previous one) is applied along the y-axis. A new line parallel to the x-axis in the slice plane is selected (each unique value of G_y determines a unique line parallel to the x-axis), which is known as phase encoding as shown in Figure 8.17.

8.2 Imaging Functional Activity

So far we have been discussing about imaging of the brain as it is in each location in the 3D space. But our ultimate objective in this chapter is to discuss on imaging the activities of the brain. Fortunately for us T_2-weighted

imaging offers an opportunity for that. Oxygenated hemoglobin in the blood and deoxygenated hemoglobin in the blood have different magnetic properties. The first one is diamagnetic (repelled by an external magnetic field) and the latter is paramagnetic (attracted by an external magnetic field). In a medium where these two different types of magnetic properties coexist (such as in the circulating blood in which oxygenated and deoxygenated hemoglobin are always present in different proportions, depending on the metabolic process it is participating in) the T_2 decay time becomes shorter. In fact this new decay time is known by a different name, T_2^*. T_2 and T_2^* are related by the following equation

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'}, \tag{8.3}$$

where T_2' is another time constant (Buxton, 2009, p. 148). Equation 8.3 can be understood in the following way. Spins sitting in different magnetic fields precess at different rates and the resulting phase dispersion reduces the net signal. The apparent relaxation time T_2^* then is less than T_2. We can separate the decay of the net signal from the sample into two processes namely, the intrinsic decay of the local signal from a small, uniform region, which is governed by T_2 and the mutual cancellation of signals from different nearby locations from the phase dispersion caused by field inhomogeneities. The net decay is described T_2^*, but it is sometimes useful to isolate the decay caused by inhomogeneities from that resulting from T_2 alone. This additional decay caused by inhomogeneities alone has been called T_2' decay. The assumed relationship among these three quantities is expressed in Equation 8.3. The inverse relaxation time is the *relaxation rate constant* and it is the rates that add to give the net relaxation effects. Any imaging modality that depicts the T_2^* contrast effects will be mapping the oxygen metabolism in the brain. Oxygen metabolism takes place due to task-related or functional activities in the brain. Therefore T_2^*-weighted imaging is known as functional magnetic resonance imaging or fMRI.

8.3 The BOLD Effect

fMRI can be generated in many different ways (see Buxton, 2009). Here we will only discuss about the *blood oxygen level dependent* (BOLD) fMRI generation, which is by far the most popular method and also does not require injecting any tracer. In a sense it is a truly noninvasive method. As shown in Figure 8.18, the demand of oxygen goes up during functional activation of a region of the brain (as small as 1 mm^3) leading to enhanced concentration

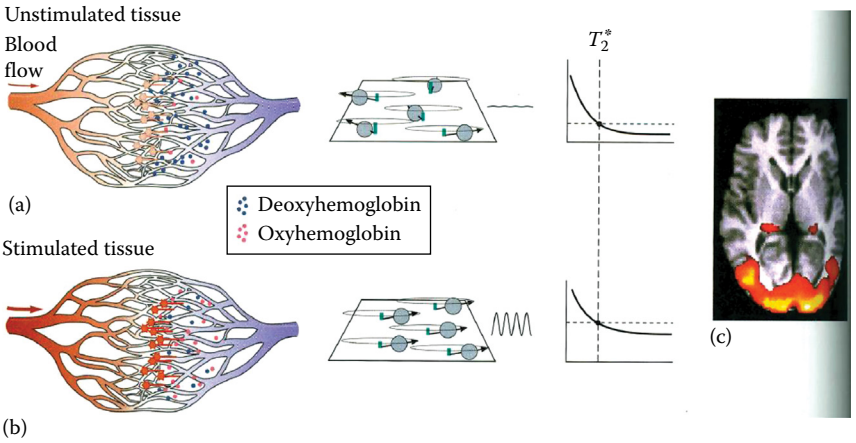


FIGURE 8.18
The region of the brain that is stimulated (active) (b) demands more oxygen than the one that is not stimulated (inactive) (a) and therefore has a greater concentration of oxyhemoglobin (b). Phase dispersion in the region will vary accordingly (middle left) leading to differential T_2^* decay time (middle right). fMRI (c) is generated by exploiting this difference in decay time. (Taken from Kandel, E.R. et al. (eds.), *Principles of Neural Science*, 4th ed., McGraw Hill, New York, 2000.)

of oxygenated hemoglobin in the blood flowing in that region. This happens because of supply of oxygenated blood is greater than its consumption in the active region. So, in the active (stimulated) brain regions the ratio of oxygenated hemoglobin to deoxygenated hemoglobin in the blood becomes higher than the regions that are not so active. On the contrary, in the unstimulated or inactive brain region the information processing is at the minimal stage. There is little activation of neurons, blood flow is not increased, and a relatively large proportion of hemoglobin is in the deoxygenated form. Because deoxygenated hemoglobin promotes efficient dephasing of the rotating protons, the T_2^* curve that characterizes fMRI is relatively steep (Figure 8.18), and the magnetic resonance signal is weak. This contrast in T_2^*-weighted imaging between the activated and inactivated regions of the brain gives rise to the BOLD fMRI images (Figure 8.18) of functional activation of the brain.

8.4 Interpreting the BOLD Response

According to the neural doctrine the activity of the brain is due to firing of the neurons. The question is how far these neuronal activities are captured in the BOLD fMRI response. Neuronal firing gives rise to brain electrophysiological

signals at different scales. The most fundamental of them are the neuronal spike trains, that is, trains of action potentials fired by one single neuron. These spike trains in turn produce a postsynaptic potential at the output synapses of the neuron. In Chapter 1 we have seen how these postsynaptic potentials travel through layered tissues to give rise to the scalp EEG. Also, when these postsynaptic potentials are collected from the surface of the cortex by touching intracranial electrode tips on it, that is called an electrocorticogram or ECoG. There is yet another kind of brain electrical signal known as the local field potential or LFP. LFP is to be collected (by inserting a depth electrode into a cortical layer) from the extracellular medium, in which the action potentials from different neurons as well as slow postsynaptic potentials are arriving and superposing on each other (or adding up together) to make a unified electric field potential (which is the LFP). LFP is usually low-pass filtered with a cutoff frequency about 300 Hz.

To summarize, we get four kinds of electrical signals from the brain due to its functional activities, namely, (1) neuronal spike train, (2) LFP, (3) ECoG, and (4) EEG. (1) carries information about the activities of a single neuron. (2), (3), and (4) carry information about localized activities of the brain in decreasing order of spatial resolution. For LFP, the location where the electrical signal is coming from is very precise, which is a little more blurred in the case of ECoG and blurred to a large extent in EEG, as can be seen from the EEG forward problem in Chapter 3. So, according to the neural doctrine, LFP is the most information-rich signal of the brain (however, this is true when LFP is allowed to retain frequency bands at least up to 3000 Hz, and not just up to 300 Hz as mentioned elsewhere in this book). It is this LFP that is most significantly correlated with fMRI (Logothetis, 2002). The BOLD signal primarily measures the input and processing of neuronal information within a brain region and not the output signal transmitted to other brain regions (but see also Sirotin and Das, 2009). The BOLD signal in a region not only depends on the local neuronal activity, but also on (1) cerebral blood volume (CBV), (2) cerebral blood flow (CBF), and (3) cortical metabolic rate of oxygen ($CMRO_2$) in that region. All these three factors, even for the same subject, may vary under different physiological and pathological conditions independent of local neuronal activities. Like any other neural signal fMRI too has artifacts, which need to be removed for an appropriate processing. So, the interpretation of BOLD fMRI is a considerable statistical challenge. The fMRI data processing pipeline has been presented in Figure 8.19. One of the most popular statistical models to address this challenge is the general linear model (GLM) implemented in a MATLAB®-based free fMRI-processing software package called Statistical Parametric Mapping (SPM). Visit http://www.fil.ion.ucl.ac.uk/spm/doc/intro/ for details.

An example of a typical BOLD signal processing from a single voxel has been explained in Figure 8.20. Let the stimulus pattern be $X(t)$ (= 1 when stimulus is on and = 0 when the stimulus is off) and its hemodynamic response

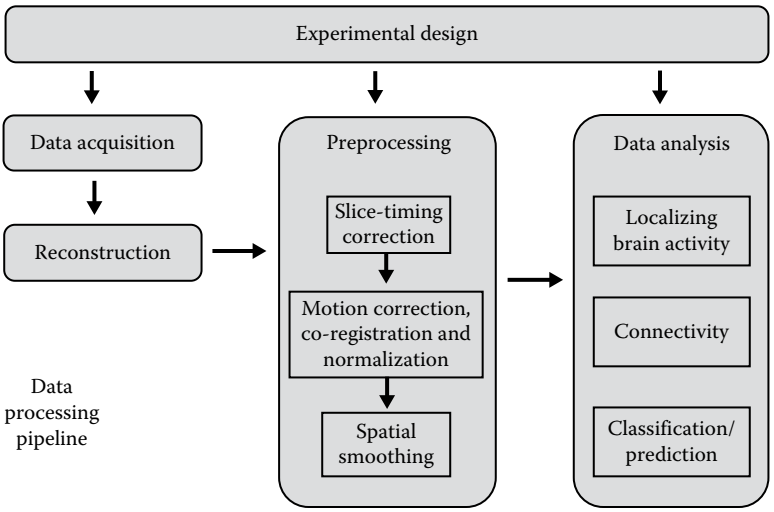


FIGURE 8.19
The fMRI data processing pipeline illustrates the different steps involved in a standard fMRI experiment. The pipeline shows the path from the initial experimental design to the acquisition and reconstruction of the data, to its preprocessing and analysis. Each step in the pipeline contains interesting mathematical and statistical problems. (Taken from Lindquist, M.A., *Stat. Sci.*, 23(4), 439, 2008.)

(or impulse response, see Chapter 2) pattern is $h(t)$. Then the expected model response to this stimulus pattern is $X(t) * h(t)$, where $*$ denotes convolution. $X(t) * h(t)$ is a smooth and delayed version of the stimulus as shown in the second line of Figure 8.20a. The activation signal (Y) is modeled as model response ($\mathbf{M}$) multiplied by an unknown amplitude a, and this BOLD response is sampled into N time points (the gap between two successive time points is called the *response time* [TR] and is associated with setting of pulse sequence at the time of fMRI acquisition). The full measured data include this sampled signal plus added noise at each of the measurement times. We assume that each noise value is independently drawn from a normal distribution with zero mean and σ standard deviation.

For this analysis, we focus on the variance of the data. Part of the variance results from the activation and so depends on the magnitude a. Because we are only interested in the variance, the first step is to remove the mean value of both the model function and the data of both the model function and the data to create a vector of model response values M_i ($i = 1$ to N) and a vector of data values Y_i. In other words, $\mathbf{M} = (M_1, \ldots, M_N)$ and $\mathbf{Y} = (Y_1, \ldots, Y_N)$, where N is the number of time points. Likewise, we can think of the noise component $\mathbf{Y}_E$ as a N dimensional vector. Clearly,

$$\mathbf{Y} = a\mathbf{M} + \mathbf{Y}_E, \tag{8.4}$$

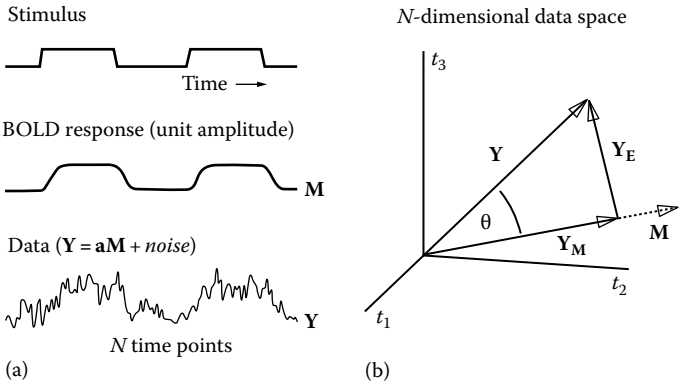


FIGURE 8.20
The general linear model analysis of BOLD data. (a) **Y** is the time series data consisting of N time points, obtained from a single voxel due to hemodynamic activity in response to stimulus + noise. The data are modeled as $\mathbf{Y} = \mathbf{aM} + noise$. (b) **Y** has been resolved into mutually perpendicular components $\mathbf{Y_M}$ (along the horizontal plane) and $\mathbf{Y_E}$ (perpendicular to the horizontal plane), signifying the actual data and the error or noise, respectively. The t-statistics, correlation coefficient r, estimated amplitude vector **a**, and noise standard deviation σ have all been modeled. (Taken from Buxton, R.B., *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.)

where a is the amplitude associated with the model response vector **M** (see Figure 8.20b). The derivations shown in Figure 8.20 are very straightforward except possibly that for $\mathbf{Y}_M$. From Figure 8.19b it is clear that

$$\mathbf{Y} \cdot \mathbf{M} = YM\cos(\theta), \tag{8.5}$$

where Y and M are length of **Y** and **M**, respectively. But by Equation 8.5 and Figure 8.20b we get

$$\mathbf{Y}_M = \mathbf{Y}\cos(\theta) = \frac{\mathbf{Y} \cdot \mathbf{M}}{M}\left(\frac{\mathbf{M}}{M}\right). \tag{8.6}$$

Now, let us fit the data in a more general linear model namely, in which there are multiple model response vectors. Following are some examples:

Removal of baseline trends: During the experiment, the fMRI signal gets contaminated by physiological and machine noise, which slowly shifts the baseline of the signal. For an accurate interpretation of the signal this baseline shift has to be eliminated.

Different types of stimuli: In sophisticated BOLD experiments multiple stimuli can be used. Responses to different stimuli need to be separated out from each other. Each type of stimulus can be modeled with a different response function.

Unknown hemodynamic response: The exact shape of the hemodynamic response is unknown and is likely to vary across the brain. Rather than using a single model function, a small set of model functions can be used to describe a range of shapes.

Event-related fMRI: The hemodynamic response itself can be estimated for each voxel by treating the response at each time point after an event as a separate model function.

An fMRI image-processing algorithm will have to account for the above fMRI components and artifacts. The general linear model (GLM) is a framework within which all these are integrated and viewed as the raw image. The fundamental assumption of GLM is that all these fMRI components and artifacts are linearly superposed to generate the raw fMRI image, from which these are to be separated out. We will explain it with a two-component example, which brings all the features of the GLM and still makes geometric visualization convenient. In Figure 8.21, the mathematical form of a GLM with two components has been explained.

Instead of a single model vector (as in Figure 8.20) we now have a matrix of model vectors with the first column representing $\mathbf{M}_1$ and the second column representing $\mathbf{M}_2$, the two model vectors. If two types of stimuli are intermixed during an experimental run, then $\mathbf{M}_1$ and $\mathbf{M}_2$ could represent the separate responses to the two stimuli. That is, $\mathbf{M}_1$ is calculated by convolving the first stimulus pattern $X_1(t)$ with the hemodynamic response function, and $\mathbf{M}_2$ is the convolution of the other stimulus pattern $X_2(t)$ with the hemodynamic response. In both the cases the vectors $\mathbf{M}_1$ and $\mathbf{M}_2$ are generated by calculating $X_1(t_i) * h(t_i)$ and $X_2(t_i) * h(t_i)$, respectively, for $i \in \{1, \ldots, N\}$. The implicit assumption here is that the responses simply add, so that the response to both stimuli is the sum of the responses to each separate stimulus.

There are now two amplitudes to be estimated, so the single amplitude a (of Figure 8.20) is to be replaced by a two dimensional vector $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$. We can think of $\mathbf{a}$ as a vector in the parameter space that defines the amplitudes of

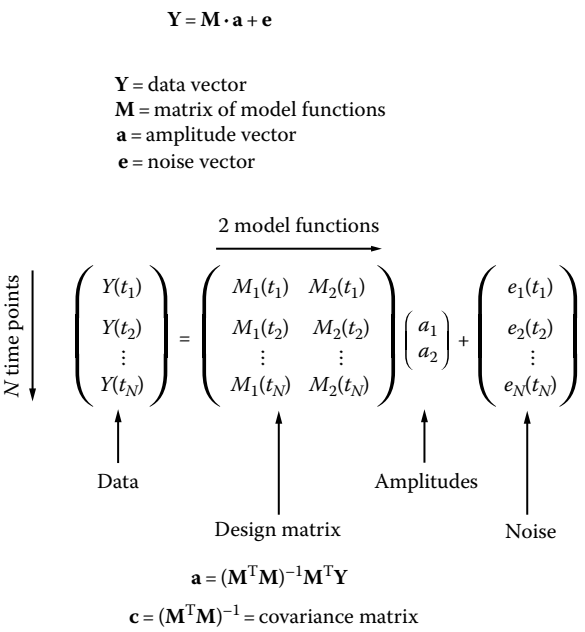


FIGURE 8.21
The mathematical structure of the general linear model. The data are modeled as a linear combination of a set of model functions with unknown amplitudes plus noise. The design matrix **M** contains one column for each model function, and the amplitudes form a vector **a**. The best-fit estimates of the amplitudes are calculated from the design matrix, as shown in the lower part of the figure. The variances of the estimated amplitudes are calculated from the covariance matrix. (Taken from Buxton, R.B., *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.)

the model functions (i.e., the vector **a** defines a point in a two-dimensional parameter space, in which one axis corresponds to a_1 and the other corresponds to a_2). But the basic form of the model is the same. The data are modeled as a sum of two model vectors with a known shape but with unknown amplitudes a_1 and a_2, plus noise with variance σ^2.

Here too we repeat the arguments of the one-model response vector case. Let us decompose the output data vector **Y** into two mutually orthogonal components as

$$\mathbf{Y} = \mathbf{Y}_M + \mathbf{Y}_E. \tag{8.7}$$

Like in Figure 8.20 here too $\mathbf{Y}_E$ is orthogonal to the space generated by $\mathbf{M}_1$ and $\mathbf{M}_2$ (i.e., $\mathbf{Y}_M$), and therefore

$$\mathbf{M}^T\mathbf{Y} = \mathbf{M}^T\mathbf{Y}_M + \mathbf{M}^T\mathbf{Y}_E = \mathbf{M}^T\mathbf{Y}_M, \tag{8.8}$$

which implies

$$\mathbf{a} = \left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T\mathbf{Y}, \tag{8.9}$$

because

$$\mathbf{Y}_M = \mathbf{M}\mathbf{a}. \tag{8.10}$$

Further details can be found in Buxton (2009, Box 15.1, p. 384). **a** is all that we need to know from the GLM, because it is the amplitude of the model response. We need to threshold the amplitude specific to a model response in order to determine if that activity or model response (in that particular voxel, the entire process described above is specific to a voxel) is significant. For an excellent exposition on GLM, see also Monti (2011).

All these steps of GLM-based fMRI processing have been taken care of in a GUI-based, freely available, popular software package, namely, *Statistical Parametric Mapping* (SPM). Although SPM is free it is based on MATLAB, which is not free. To run SPM one must have MATLAB preinstalled in the computer. Details are available at the website of Statistical Parametric Mapping, 2014. It remains the most popular software package for fMRI analysis. Since it is written in MATLAB, it is relatively readable, which makes it easy to look at the code and see exactly what is being done by the programs. Even if one does not use SPM as a primary analysis package, many of the MATLAB functions in the SPM package are useful for processing data, reading and writing data files, and other functions. SPM is also extensible through its toolbox functionality, and a large number of extensions are available via the SPM website. One unique feature of SPM is its connectivity modeling tools, including psychophysiological interaction and dynamic causal modeling. The visualization tools available with SPM are relatively limited, and many users take advantage of other packages for visualization (Poldrack et al., 2011).

One such open source software package for fMRI processing and visualization is UNIX platform–based *FMRIB Software Library* (FSL). It includes a powerful visualization tool called FSLView, which includes the ability to overlay a number of probabilistic atlases and to view the time series as a movie. It has also implemented the GLM. FSL has gained substantial popularity in recent years, due to its implementation of a number of cutting-edge techniques. First, FSL has been at the forefront of statistical modeling for fMRI data, developing and implementing a number of novel modeling, estimation, and inference techniques that are implemented in their FEAT, FLAME, and RANDOMISE modules. Second, FSL includes a robust toolbox for independent component analysis (for ICA see Section 2.2.2), which has become very popular both for artifact detection and for modeling of resting-state fMRI data. Third, FSL includes a sophisticated set of tools for analysis

of diffusion tensor imaging data, which is used to analyze the structure of white matter. Another major advantage of FSL is its integration with grid computing, which allows for the use of computing clusters to greatly speed the analysis of very large datasets (Poldrack et al., 2011).

UNIX-based open-source software *Analysis of Functional NeuroImages* (AFNI) too has a very powerful and flexible visualization abilities, including the ability to integrate visualization of volumes and cortical surfaces using the SUMA toolbox. AFNI's statistical modeling and inference tools have historically been less sophisticated than those available in SPM and FSL. However, recent work has integrated AFNI with the R statistical package, which allows the use of more sophisticated modeling techniques available within R (Poldrack et al., 2011).

Mac OS X–based Brain Voyager is yet another software for fMRI processing, but unlike the ones mentioned above, it is a proprietary product. It has an excellent user interface and is known for its user friendliness.

References

Buxton, R. B., *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.

Kandel, E. R., J. H. Schwartz, and T. M. Jessell (eds.), *Principles of Neural Science*, 4th ed., McGraw Hill, New York, 2000.

Kandel, E. R., J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth (eds.), *Principles of Neural Science*, 5th ed., McGraw Hill, New York, 2013.

Lindquist, M. A., The statistical analysis of fMRI data, *Stat. Sci.*, **23(4)**: 439–464, 2008.

Logothetis, N. K., The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal, *Philos. Trans. R. Soc. B: Biol. Sci.*, **357**: 1003–1037, 2002.

Monti, M. M., Statistical analysis of fMRI time-series: A critical review of the GLM approach, *Front. Hum. Neurosci.*, **5**: 28, 2011 available at http://journal.frontiersin.org/article/10.3389/fnhum.2011.00028/full.

Noll, D. C., A primer on MRI and functional MRI, version 2.1, Department of Biomedical Engineering and Radiology, University of Michigan, Ann Arbor, MI, 2001, available at http://www.cs.ucsb.edu/~mturk/imaging/Misc/MRI%20primer.pdf (accessed February 28, 2016).

Poldrack, R. A., J. A. Mumford, and T. E. Nichols, *Handbook of Functional MRI Data Analysis*, Cambridge University Press, New York, 2011.

Pooley, R. A., AAPM/RSNA physics tutorial for residents: Fundamental physics of MR imaging, *RadioGraphics*, **25(4)**: 1087–1099, 2005.

Sirotin, Y. B. and A. Das, Anticipatory hemodynamic signals in sensory cortex not predicted by local neuronal activity, *Nature*, **457**: 475–479, 2009.

9

Simultaneous EEG and fMRI

As we have seen in the last chapter, functional magnetic resonance imaging (fMRI) has excellent spatial localization capability; this is, however, not that precise for scalp electroencephalogram (EEG), as seen in Chapter 3. Cortical source localization of scalp EEG is performed under various constraints, according to which the accuracy greatly varies, whereas spatial localization of blood oxygenation level–dependent (BOLD) fMRI activities is relatively free of constraint. The only major constraint is that the patient should not move inside the MRI gantry during the recording. Like scalp EEG, an fMRI image too may contain a number of artifacts, which we didn't discuss in detail in the last chapter, but will mention briefly in this chapter. These artifacts may of course affect the quality of the spatial localization of BOLD activities in the fMR images.

EEG, on the other hand, which functions with an electrical signal, is very fast to reach from its source in the cortex to the scalp surface. It does so within a millisecond. Moreover, brain functions should be able to perform tasks with millisecond resolution. So EEG carries their signature to the scalp instantaneously. On the other hand, fMRI, which is dependent on the metabolic rate of glucose and the consequent oxygen requirement by the particular brain region, cannot be acquisitioned at millisecond resolution. In fact, the time resolution of the fastest fMRI acquisition technique, *echo planar imaging*, is only on the order of tens of milliseconds (Poustchi-Amin et al., 2001). So, it clubs together many minute details of brain activities at the neuronal level into one single fMRI snapshot (several such snapshots make up the fMRI video of functional activities of the brain over a given time interval). In other words, fMRI has poor temporal resolution.

Since different brain regions are specialized in performing different tasks, knowing the exact location of the brain activity almost instantaneously is advantageous. From this point of view, clubbing together EEG and fMRI is an attractive choice. (It is important to note that some EEG correlates of cognitive processing may not result in measurable changes of the BOLD signal, whereas some other patterns of neurocognitive activity may be detectable with fMRI but not with EEG (Herrmann and Debener, 2008).) But there are many formidable challenges. Perhaps the most obvious of them is how to acquire as faint an electrical signal as EEG in such a high magnetic field (order of Tesla) required for fMRI. Fortunately, many of these challenges

have been overcome to make simultaneous EEG and fMRI acquisition and analysis possible. It has become a prolific area of research since around 2000, although the first EEG acquisition during fMRI recording was made in 1993 (Ives et al., 1993). In this last chapter, we will present an overview.

9.1 Artifacts

We will use the words artifact and noise interchangeably. The difference between the two is more literary than scientific. Since localization of activities in specific brain regions is a major goal in fMRI analysis, aberration in the image due to movement of the head during image acquisition is considered a major source of artifact. Not just head movement but virtually any motion in the body may cause disturbance in the MR image. This includes coughing and swallowing, physiological motions such as pulsating blood flow, motion due to respiration, and cerebrospinal fluid (CSF) motion. All these noises are called *physiological noise*. There is also noise due to the Brownian motion of ions within the body that interferes with the magnetic field, called *thermal noise*.

We have seen in the last chapter that the fMR signal itself depends on the magnitude of the current created in a detector coil by the locally precessing magnetization in the body (Figure 8.12). The noise comes from all other sources that produce stray currents in the detector coil, such as fluctuating magnetic fields arising from random ionic currents in the body and thermal fluctuations in the detector coil itself.

The central question that must be addressed when interpreting fMRI data is whether an observed change is real, in the sense of being caused by brain activation, or whether it is a random fluctuation caused by the noise in the images. This remains a difficult problem in fMRI because the noise has several sources and is not well characterized (Buxton, 2009). Noise is a general term that describes any process that causes the measured signal to fluctuate in addition to the intrinsic nuclear magnetic resonance (NMR) signal of interest. The above-mentioned thermal noise and physiological noise are the two primary sources of noise in MRI.

The precession rates of the hydrogen nucleus in fat and water are slightly different, which causes distortion in the MRI at the boundary between fatty tissue (white matter) and water (CSF). A very common form of artifact during the fast image acquisition method, called echo planar imaging, is ghosting artifact. Due to this artifact, a shadow-like image of the original is formed either on the side, top, or bottom of the original image (depending on the phase-encoding direction). Multiple such shadows may occur one after another and two successive ones may be overlapped (see Poldrack et al., 2011 for more detail).

We discussed artifacts in EEG in Chapter 2. Acquisition of EEG under the MRI environment induces additional artifacts into the EEG. For example, the time-varying magnetic field (**B**) induces an electromotive force (EMF) in a wire loop perpendicular to the magnetic field direction. By Lenz's law, this EMF is proportional to the cross-sectional area of the wire loop and to the rate of change of the perpendicular magnetic field $d\mathbf{B}/dt$. When EEG leads are placed inside the MR scanner, the rapidly changing magnetic gradient fields and the radiofrequency (RF) pulses required for MRI may induce voltages strong enough to obscure the EEG signal. This EMF-induced current can heat the electrodes and leads, and potentially cause burn injury on the skin of the patient. Even a tiny scratching motion of the EEG leads themselves within the static field of the magnet ($\mathbf{B}_0$) inducing an EMF. Pulsating motion, related to heartbeat, yields ballistocardiographic (BCG) artifact in the EEG that can be roughly of the same magnitude as the EEG signals themselves. Further, introduction of EEG equipment into the scanner can potentially disturb the homogeneity of the magnetic field and distort the resulting MR images (Goldman et al., 2000).

EEG acquisitioned in the MR environment is subjected to two major artifacts: the fMRI gradient artifact and the BCG artifact.

9.1.1 fMRI Gradient Artifact

During the fMRI, the magnetic field inside the fMRI scanner continuously changes as a result of the switching of the magnetic field gradients. The gradients change according to the imaging sequence being used, resulting in artifacts that repeat with the collection of each new fMRI slice. The amplitude of such artifacts can be 100 times greater than the EEG signal and its frequency content overlaps that of the EEG; thus, gradient artifacts cannot be simply filtered out. In fact, it has been observed that most EEG events are in the range of 10–250 μV, whereas the amplitude of the gradient artifact in EEG signals is about 4000 μV (Allen et al., 2000). The artifact shape and amplitude vary from one EEG channel to another, depending on the location of the electrodes and the wire connections (Figure 9.1). So, the fMRI gradient artifacts cannot be filtered out (Niazy et al., 2005).

One of the earliest elementary methods for fMRI gradient artifact removal (Benar et al., 2003) consisted in selecting a few 10 s EEG pages that were recorded inside the magnet but without any scanning artifact, and using them as a baseline. Ideally, these pages should contain epileptic activity, which is the signal that needs to be preserved. There should not be any movement-related artifacts in this baseline signal. Then the frequency spectrum of this baseline was obtained. In order to filter a page with fMRI gradient artifacts, the fast Fourier transform (FFT) (see Appendix A for a discussion on FFT) of those pages had to be computed and frequencies for which the FFT amplitude differed from the baseline spectrum by more than a certain factor (typically 2 or 3) were to be noted. These frequencies were presumed to correspond to

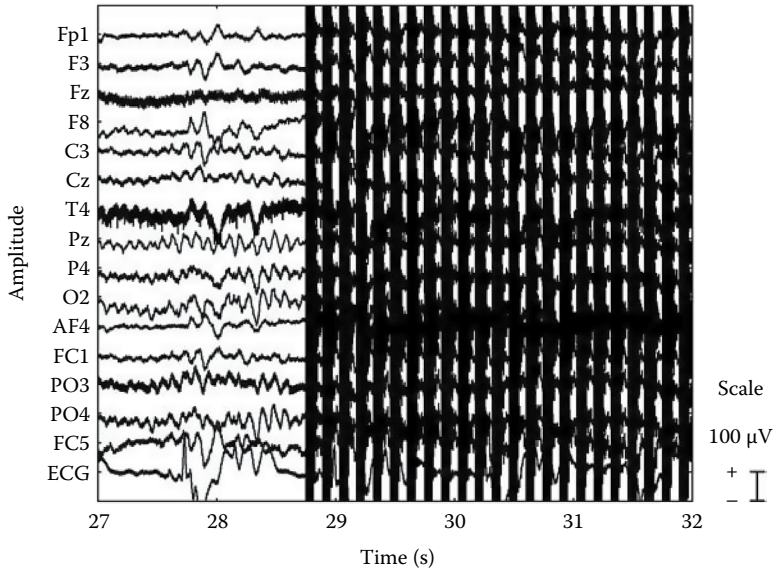


FIGURE 9.1
EEG acquisitioned in the fMR environment contaminated with fMRI gradient artifact appearing as thick vertical black stripes a little ahead of the 29th second of recording. This happens when the magnetic field changes. (Taken from Niazy, P.K. et al., *NeuroImage*, 28, 720, 2005.)

the fundamental harmonics of the fMRI gradient artifacts. Their corresponding Fourier coefficients were set to zero and an inverse Fourier transformation (see Appendix A) was obtained, yielding a signal from which the fMRI gradient artifacts should have been removed. However, this approach suffers from the typical *ringing effect* common to such frequency domain filters. The ringing effect is also known as *Gibbs phenomenon* and is due to the elimination of Fourier coefficients. The most widely used method is average artifact subtraction (Allen et al., 2000; Benar et al., 2003).

Allen and coworkers have described a two-stage gradient artifact reduction (GAR) (Allen et al., 2000), which is the most widely used method for GAR (Gotman et al., 2006). In the first stage, an average gradient artifact waveform is calculated over a fixed number of epochs and is then subtracted from the EEG for each epoch. Adaptive noise cancellation (ANC) is then done to attenuate any residual artifact (Figure 9.2).

ANC works as follows. A signal s is transmitted over a channel to a sensor, which also receives a noise n_0 uncorrelated with the signal. The combined signal and noise $s+n_0$ form the primary input to the canceller (the primary signal in Figure 9.2). A second sensor receives a noise n_1 uncorrelated with the signal but correlated in some unknown way with the noise n_0. This sensor provides the reference input to the canceller (reference signal in Figure 9.2).

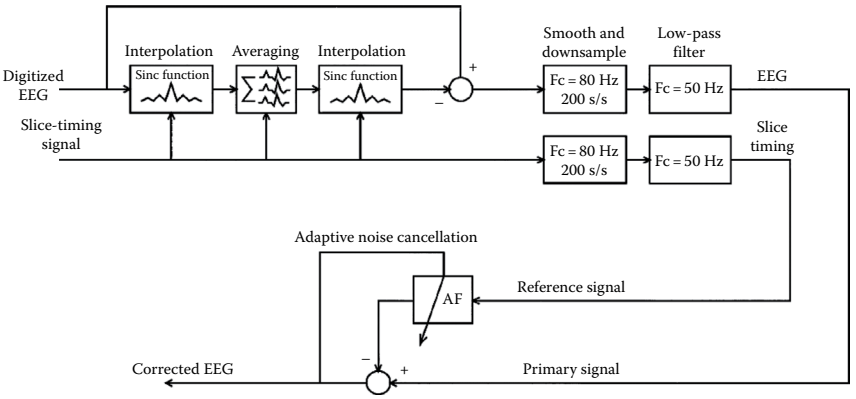


FIGURE 9.2
A schematic representation of the GAR algorithm. AF = adaptive filter, s/s = samples/second. 80 Hz low-pass filtering is for smoothing before down-sampling. 50 Hz low-pass filtering is for removing clinically irrelevant EEG components. (Taken from Allen, P.J. et al., *NeuroImage*, 12, 230, 2000.)

The noise n_1 is filtered to produce an output y that is as close a replica as possible of n_0. This output is subtracted from the primary input $s+n_0$ to produce the system output

$$z = s + n_0 - y. \tag{9.1}$$

The error signal used in an adaptive process depends on the nature of the application. In noise cancelling systems, the practical objective is to produce a system output given by (9.1) that is a best fit in the least mean squares (LMS) sense to the original signal s. This objective is accomplished by feeding the system output back to the adaptive filter and adjusting the filter through an LMS-adaptive algorithm to minimize total system output power. In other words, in an ANC system, the system output serves as the error signal for the adaptive process (Widrow et al., 1975). (Also see Chen et al., 1989 for an implementation of ANC for a biomedical time domain electrical signal like EEG.)

This can be generalized to the case where n_0 and n_1 are additive random noise components uncorrelated with each other and also with s. They can also be extended to the case where n_0 and n_1 are deterministic rather than stochastic. How to extract s from Equation 9.1 by the LMS algorithm has been very lucidly explained in a classic review by Widrow et al. (1975, Appendix A). It is technically not hard to follow, but a bit more elaborate and therefore omitted here to avoid disrupting the flow of this chapter.

On the upper left part of Figure 9.2, artifact template selection has been done, which is to be performed before the ANC. First, the digital EEG signal

is to be interpolated to transform it to an analog signal according to the following formula by Whittaker and Shannon:

$$x(t) = \sum_{n=-\infty}^{\infty} x[n] \operatorname{sinc}\left(\frac{t - nT}{T}\right), \tag{9.2}$$

where $T = 1/f_s$ and f_s is the sample frequency, $\operatorname{sinc}(x) = \sin(\pi x)/\pi x$ (the *normalized sinc function*) and $x[n]$ is the digital signal. This newly obtained analog signal $x(t)$ is then sampled at a higher rate than the original digital EEG signal $x[n]$ in order to achieve better alignment of the EEG system clock with the MR system clock.

For a periodic fMRI sequence with delays between volume scans, an epoch can be defined as the volume response or repeat time, TR. It has been suggested by Allen et al. (2000) that the EEG recording is to be divided into sections of 25 epochs, over which averaging is to be performed. According to them, the rationale for selecting 25 epochs is that most EEG events are in the range of 10–250 microV and hence the averaging should be across a minimum of 25 epochs so that the largest likely EEG event is reduced to below the smallest. This assumes that EEG is uncorrelated among epochs, which is likely as each volume scan normally lasts at least a few seconds.

For continuous fMRI acquisition, that is, no delays between volumes, an epoch can be defined as one slice (Allen et al., 2000). As the duration of a slice is typically less than 100 ms in echo planar imaging, correlation of EEG between successive epochs is likely to be significant. Allen et al. (2000) assumed that EEG is uncorrelated over 0.3 s and, for the purposes of artifact averaging, divided the EEG into sections of 25×0.3 s $= 7.5$ s. For a continuous fMRI sequence, this will correspond to $7.5 \times 13 \approx 98$ slices.

For periodic fMRI, the averaging epoch should include a period when no changing fields are applied. This is necessary as artifact is often observed to occur after the changing fields have ceased. This post-imaging artifact has two components: an exponential part with a time constant equal to that for the AC coupling in the EEG amplifier and a low-frequency decaying oscillation, possibly due to subject movement in the static field, caused by small scanner movements during application of the gradient fields. When calculating the average artifact, the first five epochs in each section are to be included. Subsequent epochs are to be included only if the cross correlation function between the epoch and the current average exceeds 0.975. Although not optimal, this scheme is a computationally efficient method for reducing averaged artifact corruption by atypical epoch signals, for example, those in which a subject movement occurs (Allen et al., 2000).

After subtracting the average artifact (Figure 9.3), the signals are to be smoothed (by low-pass filtering), in this particular case, with a cut-off frequency of 80 Hz (upper left of Figure 9.2) applied in forward and reverse directions to give zero-phase distortion (e.g., by filtfilt command in

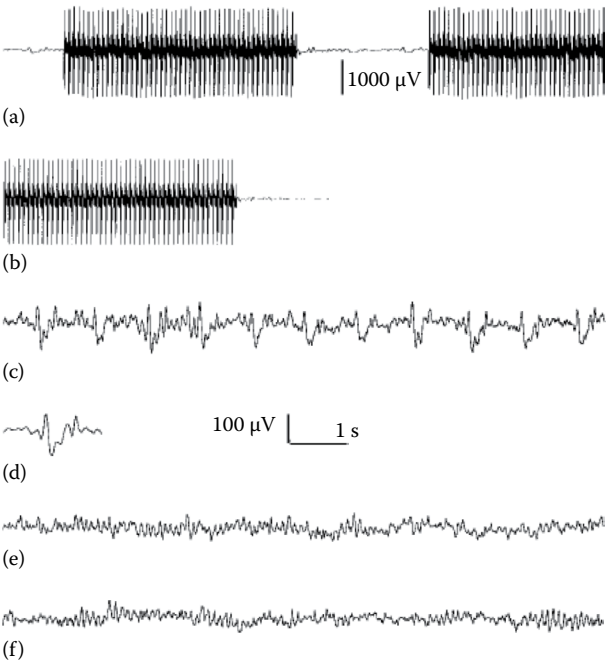


FIGURE 9.3
The EEG waveform of C4–A1 before and after the gradient and the pulse or BCG artifact (described in next section) subtraction. The bandwidth was 0.45–50 Hz. (a) The raw EEG during periodic fMRI. A large artifact is apparent during the 4 s of imaging (one volume), completely obscuring the EEG. The EEG during the 2 s gap between volume acquisitions appears to be relatively free of gradient or imaging artifact. (b) The averaged imaging artifact. Note that toward the end, the EEG is almost flat, indicating that gradient-induced patient vibration persisting after the volume acquisition has decayed and sufficient epochs have been averaged to attenuate cerebrally generated EEG. (c) The result of subtracting the averaged gradient artifact in (b) from the EEG in (a), followed by down-sampling and ANC. The BCG artifact is now clearly present and gradient artifact is difficult to discern. The display scale has been increased from that in (a) and (b). (d) The averaged BCG artifact from (c) (not to scale). (e) Result of subtracting the averaged BCG artifact in (d) from the EEG in (c). A 10.5 Hz signal is apparent in this plot and this frequency matches that of this subject's rhythm. (f) The EEG from the same subject, recorded outside the scanner, that is, free of gradient and BCG artifact. The character of this EEG appears to match closely the artifact-corrected plot in (e). (Taken from Allen, P.J. et al., *NeuroImage*, 12, 230, 2000.)

MATLAB®)) to reduce the likelihood of aliasing in the next process, which was down-sampling to 200 Hz, a sampling rate typically used for EEG. As evident in the upper left of Figure 9.3, after the average artifact template subtraction, the signal will have to be made analog again by (9.2) so as to perform the down-sampling. A low-pass filtering was then performed with a 50 Hz cutoff, assuming that EEG components above 50 Hz are not relevant for the study and also because removing high-frequency artifacts in EEG

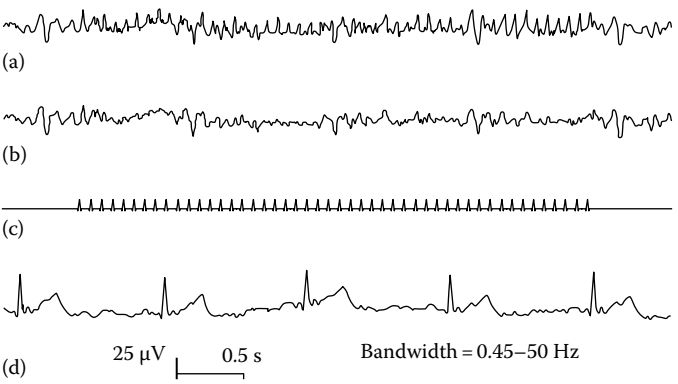


FIGURE 9.4
The effect of ANC. (a) Signal Fp2–F4 after imaging artifact subtraction. Residual artifact at the slice frequency (12.4 Hz) is clearly visible and has an amplitude of approximately 30–50 mV. (b) Same signal after ANC. (c) Reference signal used by the ANC method and has been derived digitally by low-pass filtering of the slice-timing signals. (d) Electrocardiograph (ECG) channel from the same recording and is shown here to confirm that the periodic peaks in (b) are from BCG artifact rather than an error of the subtraction method. (Taken from Allen, P.J. et al., *NeuroImage*, 12, 230, 2000.)

due to fMRI becomes difficult above this frequency. Although subtracting the averaged artifact template followed by low-pass filtering removes most of the artifact, some residual artifact, synchronized to the slice-timing signals (gradient artifact is nearly perfectly synchronized to the slice-timing signal), is sometimes observed (Figure 9.4). For the sequences used in the imaging related to Figure 9.4, this resulted in 12.4 or 13.2 Hz signals, which clearly cannot be removed by linear filtering without affecting an important part of the EEG spectra. ANC can reduce signal components correlated with a reference signal and is considered appropriate for removing this residual artifact (Allen et al., 2000).

The ANC filter (detail provided earlier in this section) can be implemented as follows (Figure 9.2). The EEG signal with averaged artifact subtracted is to be used as the primary input. The reference signal applied to the tapped delay line adaptive filter is to be generated digitally for a sample rate of 200 Hz with a value of 0 for all samples except those when slice-timing signals occur, in which case they are conventionally assigned a value of 1. This signal can then be filtered using a low-pass filter (Figure 9.4c). The adaptive filter weights can be adapted for every input sample using the Widrow–Hoff LMS algorithm to minimize error between the filter output and the primary signal (Widrow et al., 1975), as indicated earlier in this section. The entire GAR algorithm by average artifact template subtraction followed by ANC has been summarized with an implementation based on real data in Figure 9.5.

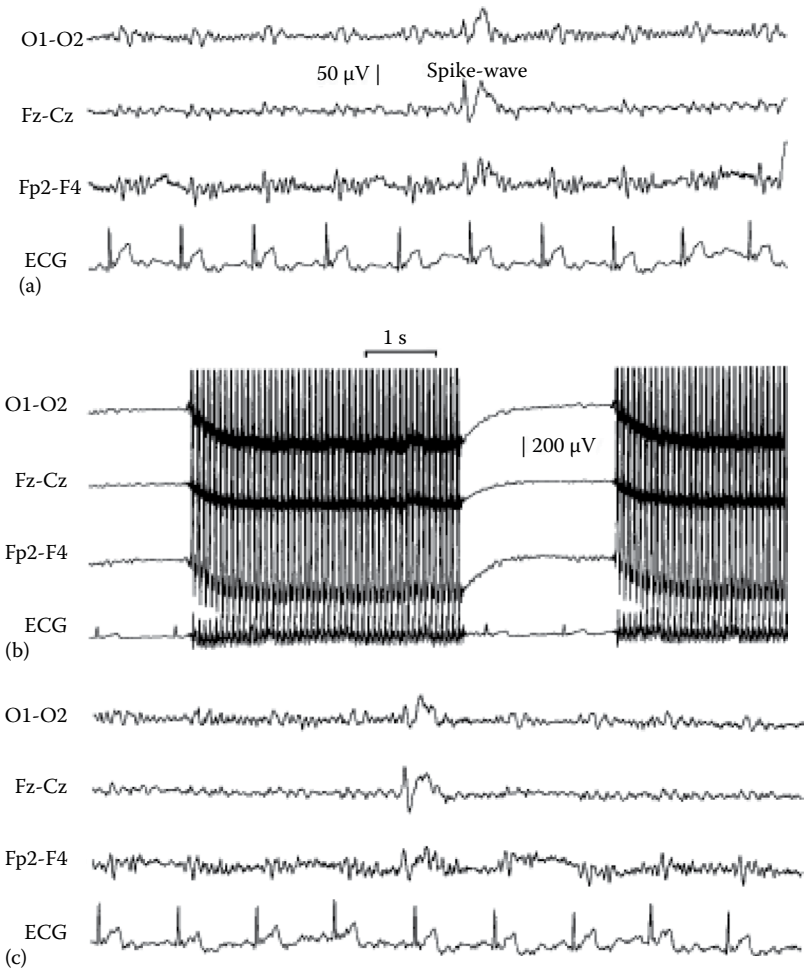


FIGURE 9.5
The EEG signals recorded inside the scanner. The ECG channel is scaled down four times lower than for the EEG plots. Bandwidth was 0.45–50 Hz. (a) These EEG signals were recorded in the absence of fMRI. A spike–wave complex added to the EEG is clearly visible, as is BCG artifact in the EEG channels (20–90 mV), synchronized to the ECG. These waveforms show the character of the EEG recorded inside the scanner, without fMRI and gradient or imaging artifact subtraction. (b) The uncorrected EEG signals (low-pass filtered and down-sampled) from the same subject during periodic fMRI. Because of the large amplitude of the artifact, all the channels have been scaled down four times compared to plots (a), (c), and (d). The large amplitude of the gradient artifact, relative to the EEG, is clearly apparent. A spike–wave complex identical to that in plot (a) has been added to the EEG, prior to GAR. This is completely obscured by the imaging artifact. (c) The effect of the GAR method on the EEG signals shown in plot (b). The spike–wave is now clearly apparent and BCG artifact synchronized to the ECG channel is also clearly visible, indicating that no significant distortion of the EEG has occurred. (*Continued*)

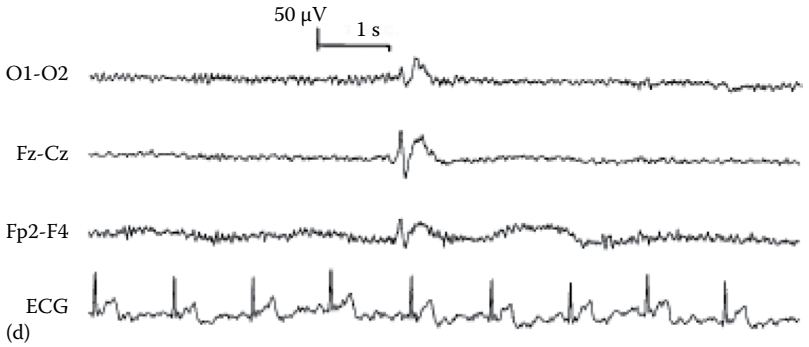


FIGURE 9.5 (*Continued*)
The EEG signals recorded inside the scanner. The ECG channel is scaled down four times lower than for the EEG plots. Bandwidth was 0.45–50 Hz. (d) BCG artifact subtraction has now been applied to the EEG signals in plot (c) and this has removed the pulse artifact almost entirely. This shows that, in this case, pulse artifact subtraction, which is vital for the interpretation of EEG recorded inside the MR scanner, can be applied effectively to signals corrected by the GAR method. (Taken from Allen, P.J. et al., *NeuroImage*, 12, 230, 2000.)

9.1.2 Cardioballistogram and Blood Flow Effect

Cardioballistogram or ballistocardiogram causes artifacts (known as *BCG* or *pulse* artifacts) in EEG due to the motion of blood through the blood vessels, which are close enough to the EEG electrodes to make them very faintly wobble around due to the motion. This motion is of course strongly dependent upon the pulsation of the heart (Figure 9.6). In the usual EEG acquisition, this does not create any noticeable effect. But inside the high magnetic field in an MRI gantry, even this negligible movement of the EEG electrodes induces an electric current flow through the electrodes, which is not related to any neural activity and therefore considered as artifact. Its magnitude is about the magnitude of the EEG and therefore causes significant signal distortion as can be seen in Figure 9.6.

BCG or pulse artifact can be removed by average artifact template subtraction (Allen et al., 1998) like in the GAR algorithm described in the last section. But in this section we will describe an algorithm for removing BCG artifacts by spatial filtering based on singular value decomposition (SVD). (For an excellent exposition on spatial filtering by SVD for EEG, see Lagerlund et al., 1997.) The raw digitized EEG is represented in an $m \times n$ matrix form $\mathbf{X}$ consisting of m time points in n channels, which is to be decomposed by SVD as

$$\mathbf{X} = \mathbf{USV}^T,$$

where $\mathbf{U}$ is an $m \times n$ matrix of m time point long signals in n channels, each of whose columns contains a signal with mean 0 and variance 1 called a

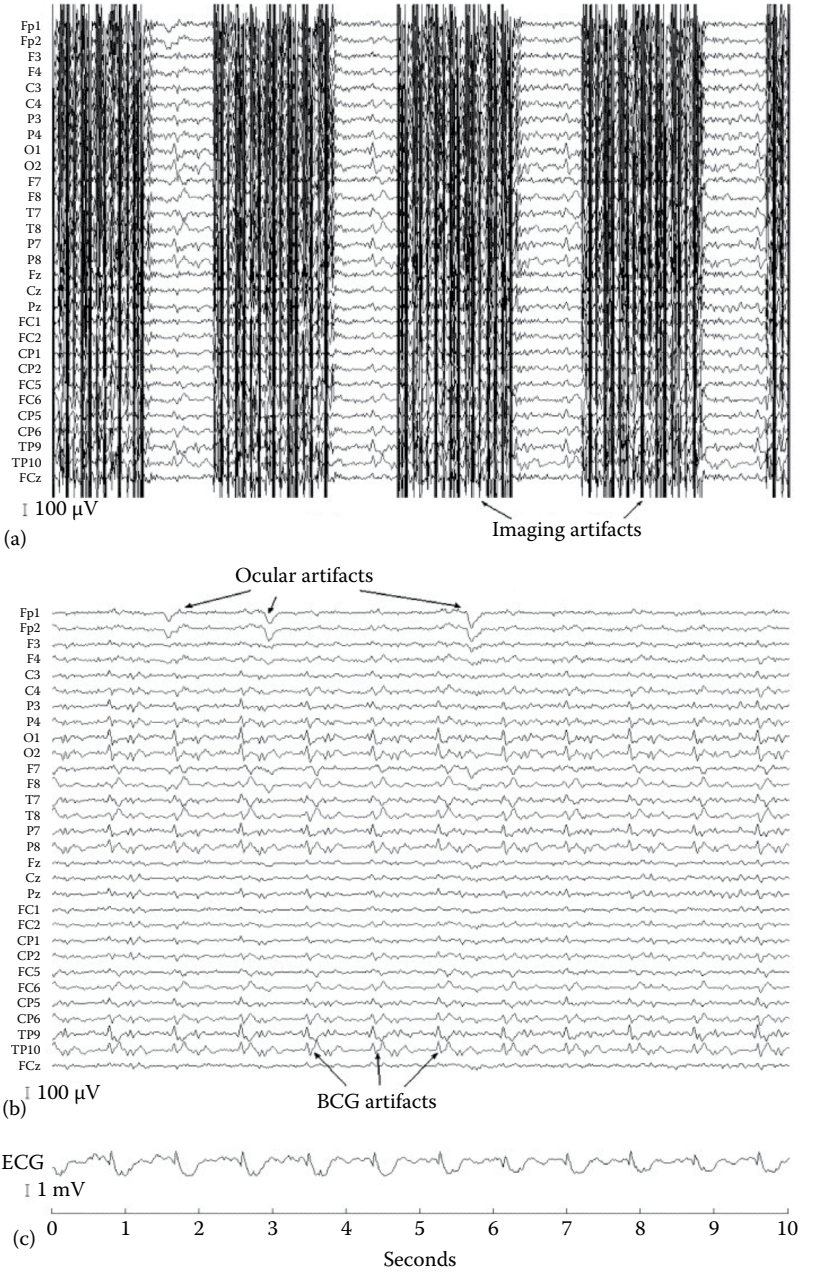


FIGURE 9.6
10 s sample of EEG acquisitioned in the MR environment. (a) The raw EEG with illustrative example of 10 s of raw EEG data collected during simultaneous EEG–fMRI, and (b) of the same traces after band-pass filtering and imaging artifact attenuation. (c) The acquired ECG signal is also shown. (Taken from Martini, D. et al., *NeuroImage*, 34(2), 598, 2007.)

normalized component signal; $\mathbf{S}$ is an $n \times n$ diagonal matrix whose diagonal entries are the standard deviation or the maximum amplitude of the raw EEG signals (whose normalized form are the columns of $\mathbf{U}$); and $\mathbf{V}$ is an $n \times n$ matrix such that $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. $\mathbf{V}$ is called *unitary matrix*, and when multiplied with a vector projects the vector from one orthonormal space to another orthonormal space.

Now let $\mathbf{W} = \mathbf{US}$, then $\mathbf{X} = \mathbf{WV}^T$. Consider the ijth entry of the matrix $\mathbf{V}$, then V_{ij} is the contribution of the ith component on the jth EEG signal. By zeroing certain columns (corresponding to certain components) of $\mathbf{V}$ we give a new matrix $\mathbf{V}'$ a new EEG epoch $\mathbf{X}' = \mathbf{WV}'^T$, which contains some but not all the components of the original EEG epoch $\mathbf{X}$. In other words, if some components can be identified as artifacts, zeroing the corresponding columns in $\mathbf{V}$ will then remove those artifacts from the new EEG epoch (Lagerlund et al., 1997).

EEG components can be visually selected with a time course corresponding to BCG activity using the ECG as a guide. Then we construct the spatial filter based on the SVD by excluding the components corresponding to the BCG in the spatial matrices. Instead of SVD, which is much akin to PCA, the component selection can be done by ICA and then be removed from the raw EEG epoch by the above method. (See Benar et al., 2002 for details.)

9.2 Recording Principles

In Chapter 8, we gave a very elementary description of fMRI acquisition, which typically takes a few to several minutes. But our brain acts on the order of milliseconds and EEG captures that almost instantaneously. In simultaneous acquisition of EEG and fMRI, the latter should be acquired at a much faster rate in order to be meaningful. Fortunately, there are techniques by which the fMRI acquisition can be made faster and one of them is *echo-planar imaging* (EPI). Most of the fMRI acquisitions are accomplished by the EPI technique. EEG sample frequency should be high (1000 Hz or more) so that a good representation of the artifacts is possible and they can then be removed effectively. This is particularly useful because the clock used in MR and EEG acquisition systems are different, often leading to sampling mismatch due to slight discrepancy between them. This is a major source of gradient-induced spikes after correction (Allen et al., 2000). Residual spikes are particularly important when the EEG sampling frequency is low. If the repetition time (TR) of the MR scanner is adjusted with the sampling rate of the EEG, this problem can be avoided (Grouiller et al., 2007).

9.2.1 EEG Wire

In an MRI acquisition environment, the wires connecting the EEG electrodes to the amplifier box are to be passed through a high magnetic field. Gradient switching that is used to alter the magnetic field gradient during the MRI acquisition can induce strong enough electromotive force (EMF) to obscure the EEG, because an EEG wire and its reference wire together form a loop. This is expressed by the formula (Allen et al., 2000)

$$V_{GA\max} = \left(dB/dt\right)_{\max} A_{\max},$$

where

$V_{GA\max}$ is the maximum amplitude of the induced gradient artifact
$(dB/dt)_{\max}$ is the maximum rate of change of the gradient field
$A_{\max}$ is the maximum loop area

In order to solve this problem, (1) bipolar montage of the EEG electrodes is to be followed and (2) for each bipolar montage electrode the wires are to be twisted together (Figure 9.7). As the wires from consecutive electrodes are twisted together, the induced EMF in the wires, both due to gradient switching and possible motions, being of the same intensity but opposite in direction, largely cancel each other.

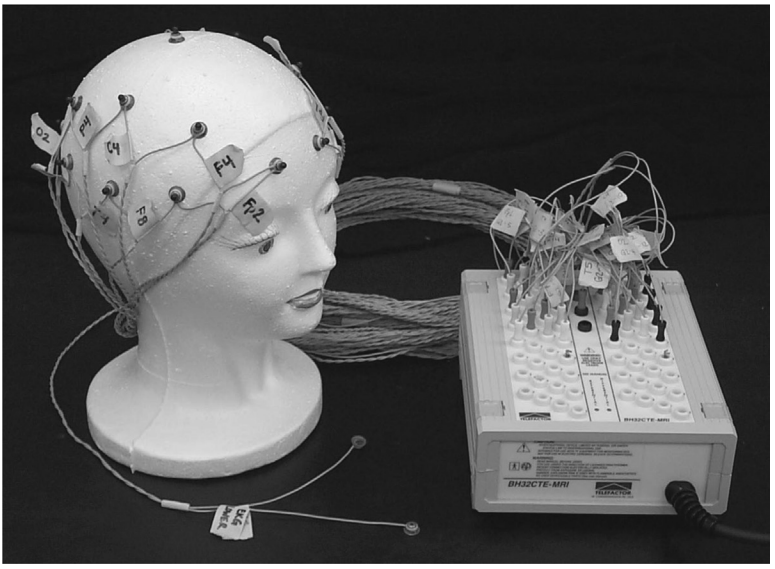


FIGURE 9.7
Bipolar montage EEG acquisition is to be followed. Leads of consecutive electrodes are twisted together to reduce gradient artifact. (Taken from Goldman, R.I. et al., *Clin. Neurophysiol.*, 111, 1974, 2000.)

EMF can also be induced due to motion in the wire in the presence of the magnetic field. In order to prevent even the slightest possible motion, all wires must be tightly secured with fixtures using bandages or adhesive tapes. Wires and electrodes must be made up of nonmagnetic materials, because magnetic materials like nickel will be strongly pulled by the main magnetic field of the machine. Wire length should be long, because the EEG amplifier will have to be outside the MRI scanner and equipment like the recording computer should be outside the scanner room.

Every fMRI is equipped with a very sensitive RF detector. If EEG equipment produces any RF wave, this would contaminate the fMRI signal. To avoid this, the EEG amplifier/digitizer should be powered by battery, not by the main AC supply.

Heat may be generated by eddy current in electrode tips under a high magnetic field. This may even cause burn injury to the skin, particularly if the tip is sharp and the field intensity is high. In order to avoid this, carbon electrodes are preferred over metal electrodes. Heat in metallic electrodes can also be reduced by adding high resistance to the electrodes.

9.2.2 Movement

As is clear from Chapter 8, MRI is sensitive to the exact positioning of the head inside the scanner and this position must be maintained throughout the scanning process, which may take as long as 45 minutes or even more. This is a difficult task even for the most cooperative subjects. The head should be rested on a cushion as comfortably as possible. This is equally important for getting good-quality EEG signal as well, because with any movement of the head the scalp electrodes and the wires attached to them will also move, producing EMF in the high magnetic field. This EMF will give rise to artifacts in the EEG.

9.3 Interpretation

Noninvasively studying the human brain by different imaging techniques has emerged as a powerful recent trend (Sui et al., 2012). The underlying premise is that one single imaging modality provides only a partial glimpse of the brain. Therefore, combining multiple imaging modalities offers us a more complete view of the structure and function of the brain than is possible by any one of the available modalities. Simultaneous EEG–fMRI studies are only a part of this paradigm, first introduced to noninvasively localize the seizure focus in a brain with focal epilepsy.

A major challenge for multimodality integration results from the fact that distinct physiological mechanisms underlie the signal generation for different

imaging modalities. Broadly speaking, the different imaging modalities can be divided into three main categories: (1) EEG/MEG (magnetoencephalography); (2) optical imaging (e.g., two-photon microscopy); and (3) fMRI/PET (positron emission tomography) (Dale and Halgren, 2001). For integration across the different modalities, a good understanding of coupling between the signals and the underlying neuronal activity is needed. Interpretation of simultaneous EEG–fMRI recording of the same cognitive task depends on the understanding of the coupling between the EEG source (Chapter 1) and the BOLD response (Chapter 8) associated with the task. Yet another challenge will be to temporally locate a transient EEG epoch of a few milliseconds duration within an fMRI snapshot of at least a few tens of milliseconds. A similar challenge will be to locate a spatially precise BOLD activation region within a larger cortical region of probable electrical activation. Three distinct approaches have been suggested to meet these challenges: (1) converging evidence, (2) direct data fusion, and (3) computational neural modeling (Horwitz and Poeppel, 2002).

9.3.1 Converging Evidence

Converging evidence is the most common method for interpreting simultaneous EEG–fMRI data. The fusion between the two different modalities is achieved by concordance in activation of EEG and BOLD signals. The best example of this is the hemodynamic response (captured by fMRI) during the interictal spike generation in the EEG of an epileptic brain. Occurrence of seizure inside the MR gantry is rare. Therefore, BOLD activation associated with an interictal spike is studied. Figure 9.8 contains a work-flow diagram of the integration of interictal spikes with BOLD response.

From Figure 9.8 we see that we first need to detect interictal spikes (Section 6.3). Next, an interictal spike train has to be built, where each spike is represented as a Dirac delta function at the time point where the interictal spike is starting. We have already seen in Chapter 8 that fMRI BOLD activation time series take distinct shapes. For example, Figure 9.9 shows shape difference in BOLD response across babies born after varied gestation periods. Standardized fMRI response (hemodynamic or BOLD response) time series is known as *fMRI karnel*. Figure 9.9 illustrates how to build an fMRI karnel from actual fMRI data. The (activation intensity) values of subsequent observations within a fixed temporal window are subjected to averaging (the tiny circle in each of the vertical bars in Figure 9.9) and the standard deviation (the half length of the vertical bar in Figure 9.9) is determined. The same is done by sliding the window, and plots as in Figure 9.9a through c are generated by fitting the best smooth curve through them (the vertical bar with circle by the LMS method, which is nothing but second-degree regression; see Hastie et al., 2009 for details). These standard templates of fMRI activation time series are preserved to be used as reference fMRI karnel templates. During a new experiment, the obtained fMRI activation time series is convolved with

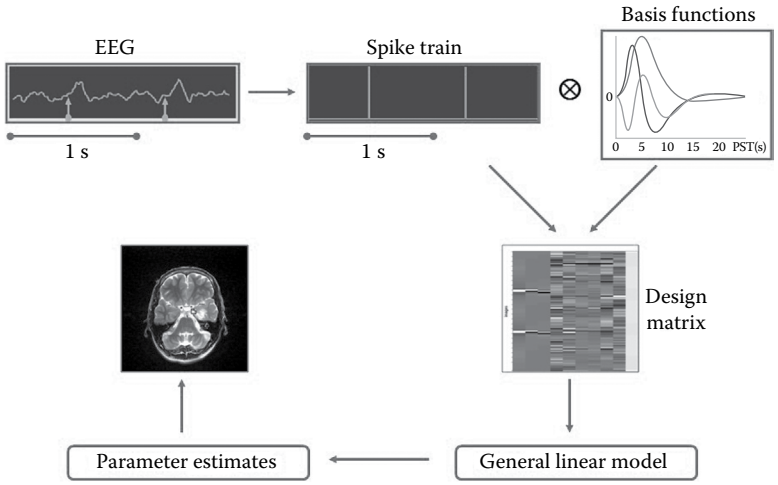


FIGURE 9.8
Interictal spike in EEG-informed study of fMRI–BOLD response. Statistical analysis of EEG–fMRI data: the epileptic interictal spikes are detected on the EEG signal; an impulse train is then created and convolved (denoted by ⊗) with a set of basic functions representing standardized BOLD activation time series. Convolved impulse trains are used as regressors of interest and are inserted into the design matrix, which is then fitted to the image data. After the estimation of the regression coefficients, inference on relevant contrasts of their estimates was performed by using a t-test or F-test depending on whether one is testing for one parameter only or several parameters at the same time. The obtained spatial t-maps or F-maps are thresholded for significance level and the resulting maps are the functional images, where the value at each voxel reflects the resemblance between the model and the data; hence the probability that this region is involved in the generation of the spikes. (Taken from Bianchi, A. et al., EEG-fMRI multimodal integration for epilepsy research, in: M.K. Gunel (ed.), *Management of Epilepsy: Research, Results and Treatment*, InTech, Rijeka, Croatia, 2011.)

an appropriate reference fMRI karnel template and the convolved signal is taken as the final fMRI activation time series. This final activation time series is a smoothed, standardized (therefore devoid of individual variations) fMRI activation.

Coming back to Figure 9.8, the EEG spike train is convolved with (judiciously chosen) standard fMRI time series templates and the convolved signal is fed into the generalized linear model (GLM) (Section 8.4) as a regressor for each voxel. For two-dimensional brain slice the image matrix that is obtained by this method is compared voxel by voxel (actually in this two-dimensional slice, voxel will reduce to pixel) with the image matrix obtained from the fMRI image of that slice. Comparison between the two image matrices is accomplished under the analysis of variance (ANOVA) (Section 4.4.1). If it is a simple comparison of differences between the two matrices (i.e., taking the values of $m \times n$ number of pixels as an $m \times n$ vector array), Student's t-test can also be performed instead of ANOVA

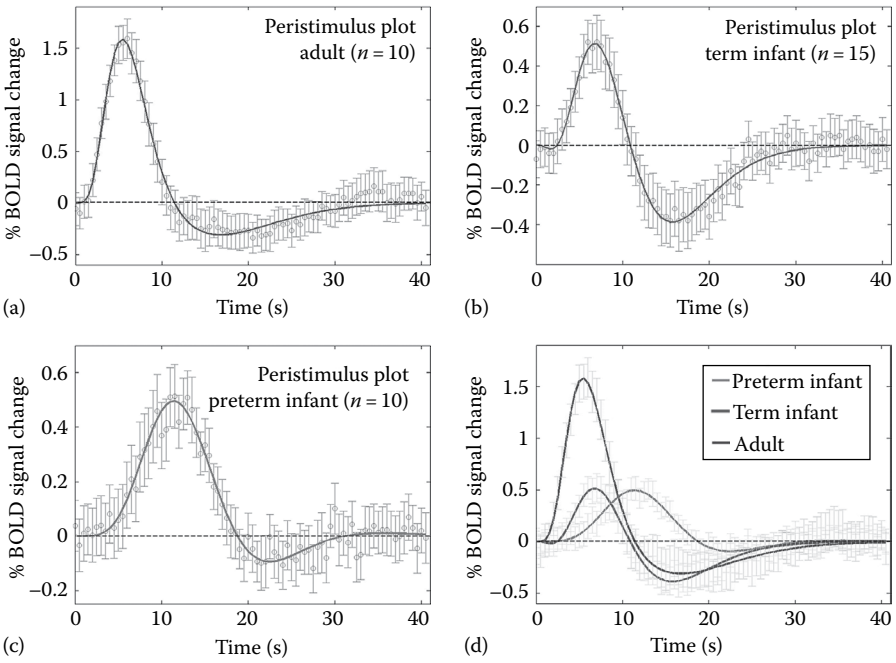


FIGURE 9.9
Peristimulus time series plots for the (a) adult, (b) term-equivalent infant, and (c) preterm infant groups. Stimulation occurred at time point 0, lasting a total of 1 s. The mean percentage BOLD signal change (relative to the prestimulus signal) at each time point (circle) is shown fitted with a double gamma probability distribution function. (d) A decrease in the time to peak of the hemodynamic response function (HRF), and an increase in peak amplitude is seen with increasing age. (Taken from Arichi, T. et al., *NeuroImage*, 63(2), 663, 2012.)

(recall ANOVA is a generalization of Student's t-test). Here the null hypothesis is that the two sets of values are from two different populations (under the assumption that each is normally distributed with the same variance). Instead of the whole slice, this test can be performed for a small region consisting of $m' \times n'$ pixels, where $m' \ll m$ and $n' \ll n$. If the null hypothesis is true with, say, probability $p < 0.001$, we claim that this particular fMRI activity is due to that particular interictal spike and locate the generator of that spike at the site of the BOLD activity of the fMRI image with a p value of 0.001, which is considered statistically significant. If the normality condition is violated and/or variances are quite different from each other, then Kruskal–Wallis test (Daniel, 1990) is to be performed instead.

One main drawback of this method is that an interictal spike in EEG may not always lead to BOLD activation. It may have BOLD deactivation instead (see Gotman et al., 2006). In addition, there is an ongoing debate regarding whether BOLD response is dependent on the total power of the neuronal

activity or on specific frequency bands of the neuronal activity (Rosa et al., 2010). This may be an important issue at the time of constructing the regressors in the GLM for fMRI analysis.

9.3.2 Direct Data Fusion

In the direct data fusion approach, two data sets are directly combined using some quantitative methods. Here the basic premise is that the EEG and fMRI signals generated by each method correspond to the same set of underlying neural generators. The most common method that has been employed to combine hemodynamic and electromagnetic data assumes that there are a few underlying equivalent current dipoles that generate the EEG and use the local maxima obtained by fMRI as constraints on localizing these EEG dipole sources (Horwitz and Poeppel, 2002).

A variety of methods have been suggested for investigating the relationship between fMRI and oscillatory EEG components. For instance, several studies on the human posterior alpha rhythm have demonstrated that the power spectrum of wavelet decomposition (not the usual Fourier power spectrum, but determined exactly in the same way by substituting the class of Fourier basis functions by a suitable class of wavelet basis functions) convolved with an fMRI time series template, as described in the previous section, can be correlated electrode- and voxel-wise with the BOLD signal (see Goldman et al., 2002 for technical details). Classical EEG rhythms can be recorded over wide scalp areas, not only because they are distributed over wide cortical areas but also due to volume conduction that spreads electric activity across the scalp. Therefore, single EEG traces do not represent electric activity from the directly underlying cortex, but are rather a weighted sum of electric activity from the entire surface of the brain. As a consequence, the broadband power of an EEG trace may reflect a mixture of functionally and topographically different EEG rhythms, prohibiting a straightforward interpretation of fMRI correlates. When the area of interest is in fMRI correlates of subtle EEG rhythms with small amplitudes, which are often masked by more prominent EEG rhythms, a clear separation of rhythms is necessary. In this case, statistical methods like principal component analysis (PCA) or independent component analysis (ICA) may be employed to estimate the number of uncorrelated independent field patterns in the raw data. Physiological EEG rhythms, however, are unlikely to be mutually orthogonal or independent. To separate rhythms by means of statistical methods while accounting for this consideration, it may be necessary to functionally modulate individual rhythm strength (Ritter and Villringer, 2006).

Another method for combining simultaneous EEG and fMRI data fusion has been outlined in Figure 9.10. This is an extension of partial least squares (PLS) regression. PLS analysis is based on the SVD of the covariance matrix between fMRI data (voxel by time matrix) and a matrix of independent variables. The extended version of multiway partial least squares regression

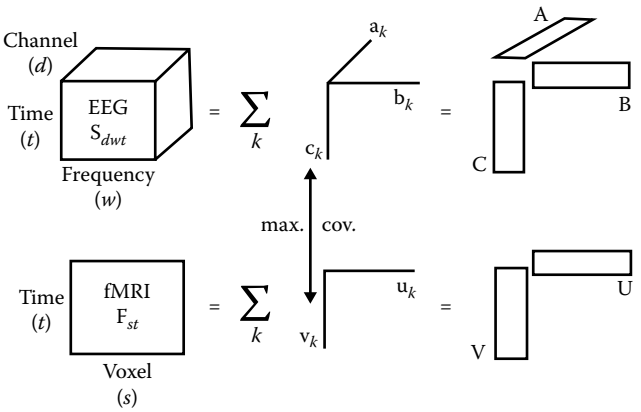


FIGURE 9.10
The time-varying EEG spectrum is represented as a three-dimensional array indexed by time t, frequency w, and channel d. The fMRI matrix is indexed by time t and voxels s. Both data are decomposed into a sum of components. Each EEG component has spatial $\mathbf{a}_k$, spectral $\mathbf{b}_k$, and temporal $\mathbf{c}_k$ signatures. fMRI components have a spatial $\mathbf{u}_k$ and a temporal $\mathbf{v}_k$ signature. The extraction of components is performed simultaneously by constraining the temporal signatures to have maximal covariance. Joining all components for each signature allows them to be expressed in matrix notation, obtaining the corresponding matrices **A**, **B**, **C**, **V**, and **U**. (Taken from Martinez-Montes, E. et al., *NeuroImage*, 22(3), 1023, 2004.)

(N-PLS) can deal with multidimensional data, decomposing EEG and fMRI data into a sum of components as shown in Figure 9.10. Each EEG component is the Euclidean product of spatial, spectral, and temporal signatures and each fMRI component is the Euclidean product of spatial and temporal signatures. The idea behind this data-driven approach is to estimate the linear combination of EEG signals that correlate optimally with the BOLD signal and a linear combination of fMRI voxel signals that correlate optimally with the EEG. Transient waveforms are not yet optimally described in the time frequency framework. However, this approach holds considerable promise for future research in this area (Ritter and Villringer, 2006).

We have discussed the cortical source localization of scalp EEG signals in Chapter 3. Some studies have demonstrated good correspondence between cortical electrical sources and cortical sites of fMRI activities (Moosmann et al., 2004). However, at times, there are likely to be obvious mismatches between the two measures. A number of situations are easily conceivable when EEG and fMRI signals may diverge. For example, fMRI signals from deep-brain structures like the hippocampus, amygdala, or ventral tegmental area (VTA), from where no electrical signals can reach up to the scalp due to the depth of those brain areas. Therefore, the amygdala, hippocampus, or VTA will have fMRI signals, but no corresponding EEG signals. On the other hand, EEG signals can be large if only a few percent of neurons are

synchronously active in selective frequency bands. However, the majority of neurons may be inactive, producing minimal metabolic activities that may be measured by fMRI (Moosmann et al., 2004).

9.3.3 Computational Neural Modeling

The third way by which EEG and fMRI data can be compared is through the use of computational neural models that can simulate the different data types. The idea here is to construct a large-scale biologically realistic neural network model that can perform the cognitive tasks under investigation. The model would be constructed so as to be able to generate simulated fMRI signals and simulated EEG signals that can be compared to experimentally observed values. The critical notion is that data types with different spatiotemporal properties are not compared directly to one another, but are compared inside a neural model that incorporates specific hypotheses about how particular cognitive operations are mediated at the neural level. That is, the assumptions one makes are about how macroscopically measured data are related to neuronal physiology, not about how these data are related to each other. The major disadvantage of this approach is that modeling is meant to simplify what actually is going on and thus it is hard to know if lack of agreement between computational and experimental results means the model and its corresponding hypotheses are too simple or just wrong (Horwitz and Poeppel, 2002).

EEG and fMRI are both generated by a large network, consisting of neurons on the order of millions spread across different cortical areas. Detailed modeling for individual neurons and then modeling the network with detailed modeling of the synaptic connections will be computationally intractable in this case. Therefore, a *neural mass modeling*, which describes the average neuronal activity with a small number of state variables, will be practical. These states summarize the behavior of millions of interacting neurons. Basically, these models use two conversion operations: a wave-to-pulse operator at the soma of neurons, which is generally a static sigmoid function, and a linear pulse-to-wave conversion implemented at a synaptic level, within the ensemble. The first operator relates the mean firing rate to average postsynaptic depolarization. This is assumed to be instantaneous. The second operator depends on synaptic kinetics and models the average postsynaptic response as a linear convolution of incoming spike rate. The shape of the convolution kernels embodies the synaptic and dendritic kinetics of the population (David and Friston, 2003).

Event-related potential (ERP) with respect to a stimulus presentation and the corresponding BOLD response have been simulated with the help of a neural mass model (Babajani and Soltanian-Zadeh, 2006). In this model, cortical minicolumns are the smallest computational units (Mountcastle, 1997). A minicolumn consists of a narrow chain of neurons extending vertically across the cellular layers II to VI. Each minicolumn in primates

contains roughly 80–100 neurons. The width of each minicolumn is 50 μm and the mean value for intercolumnar distance is 80 μm (Buxhoeveden and Casanova, 2002). There are three basic cell types in minicolumns: the stellate cells, the local inhibitory interneurons, and the pyramidal cells. The axon of the first two types spread vertically in their minicolumn without any considerable outputs to the neighboring minicolumns. The output of a minicolumn is mainly derived from its pyramidal cells and so all cell types in a minicolumn receive input from pyramidal cells of the neighboring minicolumns. The stellate cells also receive afferent thalamic input (Buxhoeveden and Casanova, 2002).

The block diagram of the neural mass model for generation of EEG and fMRI has been elaborated in Figure 9.11 with a detailed caption. Since computational neural modeling for simulating the simultaneous EEG and fMRI

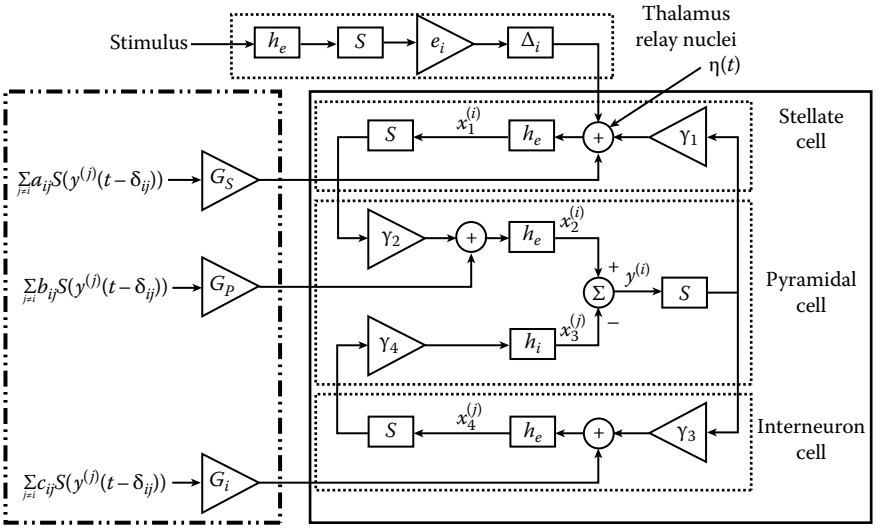


FIGURE 9.11
An illustration of the neural mass model for the *i*th minicolumn. The solid box shows the classical Jansen's model (Jansen and Rit, 1995). The left dash-dot box illustrates contributions of the neighboring minicolumns to the *i*th minicolumn. The EEG signal is related to $y^{(i)}$. h_e and h_i are impulse responses of the excitatory and the inhibitory synapses. $S(\cdot)$ is the sigmoid function. $x_k^{(i)}$ shows overall postsynaptic potential (PSP) of different cell populations. The four constants γ_i, $i \in \{1,2,3,4\}$ represent the total number of synapses in each subpopulation. G_S, G_P, and G_I represent the influence of the neighboring minicolumns on the stellate cells, pyramidal cells, and interneurons, respectively. a_{ij}, b_{ij}, and c_{ij} represent the strength of the connections of different cell populations between the *i*th and the *j*th minicolumns. δ_{ij} is the propagation delay between minicolumns *i* and *j*. c_i represents the strength of the afferent input to the *i*th minicolumn. Δ_i is the propagation delay between external stimulus ($Stim(\cdot)$) and cortical columns. $\eta(t)$ is physiological noise. (Taken from Babajani, A. and Soltanian-Zadeh, H., *IEEE Trans. Biomed. Eng.*, 53(9), 1794, 2006.)

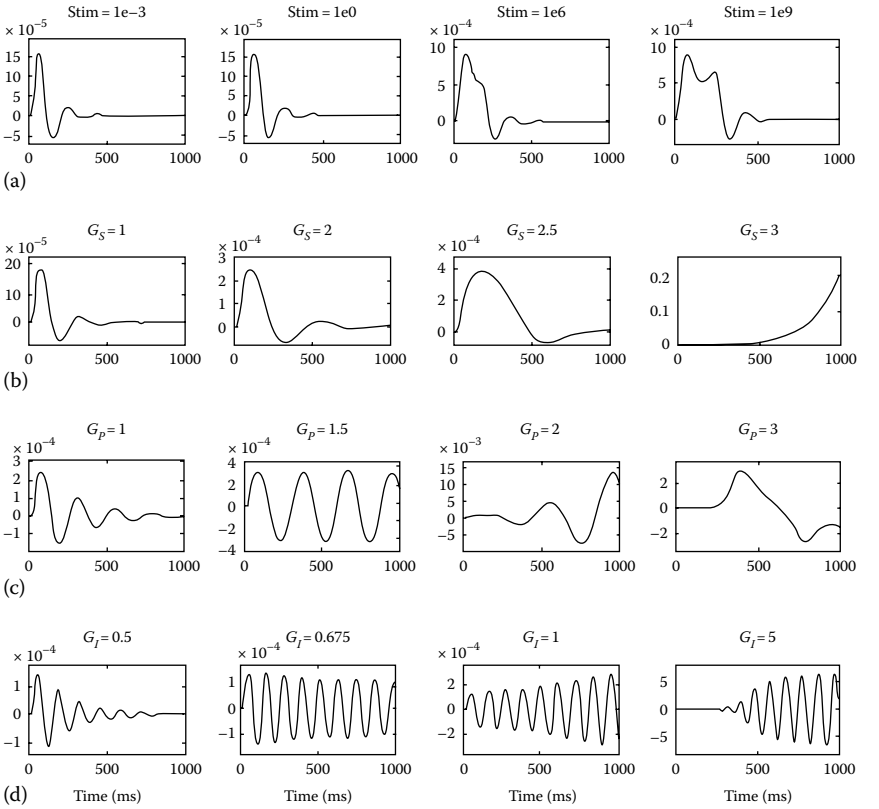


FIGURE 9.12
Illustration of the capability of the neural mass model for producing the ERPs. The stimulus is the Dirac delta function. In each subplot, all parameters are selected as the default values except for the parameters whose values are mentioned at the top of each subplot. (a) Effect of the external stimulus strength on ERP, $\sigma_E = 0$ in this simulation. (b) Effect of the stellate cells gain (G_S) on ERP. (c) Effect of the pyramidal cells gain (G_P) on ERP. (d) Effect of the inhibitory interneurons gain (G_I) on ERP. (Taken from Babajani, A. and Soltanian-Zadeh, H., *IEEE Trans. Biomed. Eng.*, 53(9), 1794, 2006.)

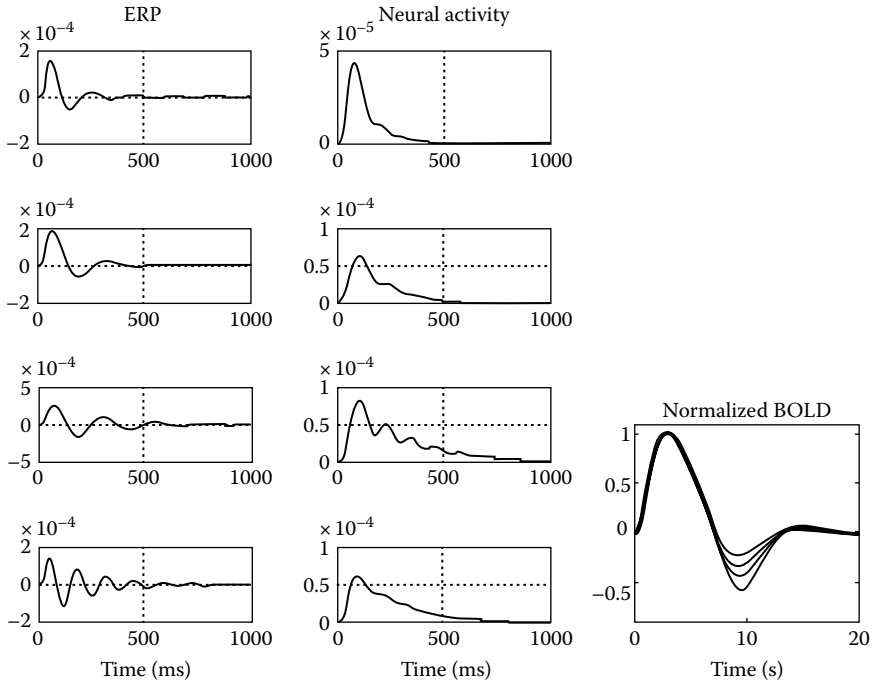


FIGURE 9.13
Illustrations of the effects of the proposed integrated model parameters on the ERP, the neural activity, and the BOLD signals. All conditions are similar to the left column in Figure 9.12. The BOLD graphs correspond to the four subplots in the middle column, respectively, where, moving from the top to the bottom subplots, their BOLD undershoots increase. (Taken from Babajani, A. and Soltanian-Zadeh, H., *IEEE Trans. Biomed. Eng.*, 53(9), 1794, 2006.)

data is not a very popular approach, we are not presenting the detailed mathematical equations here. Interested readers are encouraged to read the paper by Babajani and Soltanian-Zadeh (2006). This model has generated ERP (Figure 9.12) in response to a stimulus and the corresponding BOLD activation (Figure 9.13) in the cortex.

References

Allen, P. J., O. Josephs, and R. Turner, A method for removing imaging artifact from continuous EEG recorded during functional MRI, *NeuroImage*, **12**: 230–239, 2000.

Allen, P. J., G. Polizzi, K. Krakow, D. R. Fish, and L. Lemieux, Identification of EEG events in the MR scanner: The problem of pulse artifact and a method for its subtraction, *NeuroImage*, **8**: 229–239, 1998.

Arichi, T., G. Fagiolo, M. Varela, A. Melendez-Calderon, A. Allievi, N. Merchant, N. Tusor et al., Development of BOLD signal hemodynamic responses in the human brain, *NeuroImage*, **63(2)**: 663–673, 2012.

Babajani, A. and H. Soltanian-Zadeh, Integrated MEG/EEG and fMRI model based on neural masses, *IEEE Trans. Biomed. Eng.*, **53(9)**: 1794–1801, 2006.

Benar, C., Y. Aghakhani, Y. Wang, A. Izenberg, A. Al-Asmi, F. Dubeau, and J. Gotman, Quality of EEG in simultaneous EEG-fMRI for epilepsy, *Clin. Neurophysiol.*, **114**: 569–580, 2003.

Benar, C.-G., D. W. Gross, Y. Wang, V. Petre, B. Pike, F. Dubeau, and J. Gotman, The BOLD responses to interictal epileptiform discharges, *NeuroImage*, **17(3)**: 1182–1192, 2002.

Bianchi, A., T. Franchin, and M. G. Tana, EEG-fMRI multimodal integration for epilepsy research, in: M. K. Gunel (ed.), *Management of Epilepsy: Research, Results and Treatment*, InTech, Rijeka, Croatia, 2011.

Buxhoeveden, D. and M. F. Casanova, The minicolumn hypothesis in neuroscience: A review, *Brain*, **125**: 935–951, 2002.

Buxton, R. B., *Introduction to Functional Magnetic Resonance Imaging*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2009.

Chen, J., J. Vandewalle, W. Sansen, G. Vantrappen, and J. Janssens, Adaptive method for cancellation of respiratory artefact in electrogastric measurements, *Med. Biol. Eng. Comput.*, **27(1)**: 57–63, 1989.

Dale, A. M. and E. Halgren, Spatiotemporal mapping of brain activity by integration of multiple imaging modalities, *Curr. Opin. Neurobiol.*, **11(2)**: 202–208, 2001.

Daniel, W. W., *Applied Nonparametric Statistics*, 2nd ed., PWS-Kent Publishing Co., Boston, MA, 1990.

David, O. and K. J. Friston, A neural mass model for MEG/EEG: Coupling and neuronal dynamics, *NeuroImage*, **20**: 1743–1755, 2003.

Goldman, R. I., J. M. Stern, J. Engel, Jr., and M. S. Cohen, Acquiring simultaneous EEG and functional MRI, *Clin. Neurophysiol.*, **111**: 1974–1980, 2000.

Goldman, R. I., J. M. Stern, J. Engel, Jr., and M. S. Cohen, Simultaneous EEG and fMRI of the alpha rhythm, *Neuroreport*, **13(18)**: 2487–2492, 2002.

Gotman, J., E. Kobayashi, A. P. Bagshaw, C.-G. Benar, and F. Dubeau, Combining EEG and fMRI: A multimodal tool for epilepsy research, *J. Mag. Reson. Imaging*, **23(6)**: 906–920, 2006.

Grouiller, F., L. Vercueil, A. Krainik, C. Segebarth, P. Kahane, and O. David, A comparative study of different artifact removal algorithms for EEG signals acquired during functional MRI, *NeuroImage*, **38(1)**: 124–137, 2007.

Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Theory: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, 2009.

Herrmann, C. S. and S. Debener, Simultaneous recording of EEG and BOLD responses, *Int. J. Psychophysiol.*, **67**: 161–168, 2008.

Horwitz, B. and D. Poeppel, How can EEG/MEG and fMRI/PET data be combined? *Hum. Brain Mapp.*, **17(1)**: 1–3, 2002.

Ives, J. R., S. Warach, F. Schmitt, R. R. Edelman, and D. L. Schomer, Monitoring the patient's EEG during echo planar MRI, *Electroencephal. Clin. Neurophysiol.*, **87(6)**: 417–420, 1993.

Jansen, B. H. and V. G. Rit, Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns, *Biol. Cybern.*, **73(4)**: 357–366, 1995.

Lagerlund, T. D., F. W. Sharbrough, and N. E. Busacker, Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition, *J. Clin. Neurophysiol.*, **14(1)**: 73–82, 1997.

Martinez-Montes, E., P. A. Valdes-Sosa, F. Mirwakeichi, R. I. Goldman, and M. S. Cohen, Concurrent EEG/fMRI analysis by multiway Partial Least Squares, *NeuroImage*, **22(3)**: 1023–1034, 2004.

Martini, D., M. G. Perucci, S. Cugini, A. Ferretti, G. L. Romani, and C. Del Gratta, Complete artifact removal for EEG recorded during continuous fMRI using independent component analysis, *NeuroImage*, **34(2)**: 598–607, 2007.

Moosmann, M., P. Ritter, J. Steinbrink, and A. Villringer, Simultaneous multimodal acquisition of surface-EMG, EEG and fMRI, *Proc. Int. Soc. Mag. Reson. Med.*, **11**: 1045, 2004.

Mountcastle, V., The columnar organization of the neocortex, *Brain*, **120**: 701–722, 1997.

Niazy, P. K., C. F. Beckmann, G. D. Innetti, J. M. Brady, and S. M. Smith, Removal of FMRI environment artifacts from EEG data using optimal basis sets, *NeuroImage*, **28**: 720–737, 2005.

Poldrack, R. A., J. A. Mumford, and T. E. Nichols, *Handbook of Functional MRI Data Analysis*, Cambridge University Press, Cambridge, U.K., 2011.

Poustchi-Amin, M., S. A. Mirowitz, J. J. Brown, R. C. McKinstry, and T. Li, Principles and applications of echo-planar imaging: A review for general radiologist, *RSNA RadioGraphics*, **21(3)**: 767–779, 2001.

Ritter, P. and A. Villringer, Simultaneous EEG-fMRI, *Neurosci. Biobehav. Rev.*, **30**: 823–838, 2006.

Rosa, M. J., J. Kilner, F. Blankenburg, O. Josephs, and W. Penny, Estimating the transfer function from neuronal activity to BOLD using simultaneous EEG-fMRI, *NeuroImage*, **49(2)**: 1496–1509, 2010.

Sui, J., T. Adali, Q. Yu, and V. D. Calhoun, A review of multivariate methods for multimodal fusion of brain imaging data, *J. Neurosci. Methods*, **204(1)**: 68–81, 2012.

Widrow, B., J. R. Glover, J. M. McCool, K. Kaunitz, C. S. Williams, R. E. Hearn, J. R. Zeidler, E. Dong, and R. C. Goodlin, Adaptive noise cancelling: Principles and applications, *Proc. IEEE*, **63(12)**: 1692–1716, 1975.

Appendix A: Fourier Transformation

Time-frequency analysis is perhaps the most common signal-processing technique applied on the electroencephalogram (EEG) signals. By this analysis, we try to know which frequency component has how much strength in the signal. Two standard approaches for time-frequency analysis are Fourier transformation and wavelet transformation. In this appendix, we discuss the former and take up the latter in Appendix B.

Although Fourier transform is often applied on arbitrary signals with significant results, it is most appropriate on signals that (1) are periodic and (2) satisfy Dirichlet's conditions. A signal $x(t)$ is *periodic* with period p if and only if there exists a smallest real number $p>0$, such that $x(t+p)=x(t)$ for all t. In the signal-processing literature, an equivalent condition, called *statistically stationary*, is more popular. By statistically stationary signal $x(t)$, we mean there exists a smallest $p>0$, such that all successive nonoverlapping windows of length p, each of $x(t)$, have first-order (mean), second-order (variance), third-order (proportional to skewness), fourth-order (proportional to kurtosis), and, in general, all nth-order moments for $n\geq 1$, equal across the windows. In other words, the probability distribution of values of $x(t)$ in all windows will be the same.

A periodic signal $x(t)$ with period p satisfies Dirichlet's conditions if and only if the following conditions hold:

1. $x(t)$ has a finite number of discontinuities for $t\in[kp, (k+1)p]$, where k is any integer.
2. $x(t)$ has a finite number of maxima and minima for $t\in(kp, (k+1)p)$, where k is any integer.
3. $x(t)$ is absolutely integrable over $[kp, (k+1)p]$, that is,

$$\int_{kp}^{(k+1)p} \left|x(t)\right| dt < \infty. \tag{A.1}$$

All periodic signals of practical interest satisfy these conditions.

If $x(t)$ satisfies conditions (1) with period p and (2), it can be represented by its Fourier series expression

$$x(t) = \sum_{n=-\infty}^{\infty} a_n \cos\left(\frac{2\pi n}{p}t\right) + \sum_{n=-\infty}^{\infty} b_n \sin\left(\frac{2\pi n}{p}t\right), \tag{A.2}$$

for each real t. In other words, if conditions (1) and (2) are satisfied, no matter how complicated the signal looks, it can always be expressed as a weighted sum of sine and cosine signals. Sinusoidal signals are among the simplest-looking signals.

The nth continuous Fourier transform $F_n(x(t))$ of $x(t)$ is defined as follows:

$$F_n\left(x(t)\right) = \int_{-p/2}^{p/2} x(t)\exp\left(-i\frac{2\pi n}{p}t\right)dt, \tag{A.3}$$

where n is any integer and $i = \sqrt{-1}$. The discrete form of this transformation is

$$F_n\left(x(m)\right) = \sum_{m=-\infty}^{\infty} x(m)\exp\left(-i\frac{2\pi n}{p}m\right). \tag{A.4}$$

Notice that

$$\cos\left(\frac{2\pi n}{p}t\right) = \frac{\exp\left(i\frac{2\pi n}{p}t\right) + \exp\left(-i\frac{2\pi n}{p}t\right)}{2} \tag{A.5}$$

and

$$\sin\left(\frac{2\pi n}{p}t\right) = \frac{\exp\left(i\frac{2\pi n}{p}t\right) - \exp\left(-i\frac{2\pi n}{p}t\right)}{2i}. \tag{A.6}$$

Substituting (A.5) and (A.6) in (A.2), we can write $x(t)$ as a complex exponential series on which the nth Fourier transform (A.3) or (A.4) is to be performed. By elementary integral calculus, it is straightforward to show that the nth Fourier transform of $x(t)$ produces the complex number $a_n + ib_n$, where $x(t)$ satisfies (1) with period p and (2), and therefore can be expressed as in (A.2). To be more precise,

$$a_n = \frac{1}{p}\int_{-p/2}^{p/2} x(t)\cos\left(\frac{2\pi nt}{p}\right)dt \quad \text{and} \quad b_n = \frac{1}{p}\int_{-p/2}^{p/2} x(t)\sin\left(\frac{2\pi nt}{p}\right)dt. \tag{A.7}$$

For a derivation of (A.7) see subsection 4.1.1 of Proakis and Manolakis (2007). We call $a_n^2 + b_n^2$ the *power* associated with the nth frequency component

of the signal $x(t)$, where $a_n \cos\left(\frac{2\pi n}{p}t\right) + b_n \sin\left(\frac{2\pi n}{p}t\right)$ is the nth harmonic or the nth frequency component of $x(t)$. The sequence $\{a_n + ib_n\}_{n=-\infty}^{\infty}$ constitutes the so-called frequency domain representation of the time domain signal $x(t)$. The sequence $\{a_n^2 + b_n^2\}_{n=-\infty}^{\infty}$ gives the *power spectrum* of $x(t)$. Typically, the sequence $\{a_n^2 + b_n^2\}_{n=-\infty}^{\infty}$ convoluted with a window function (e.g., Welch) gives the typical power spectrum plots for $x(t)$.

In reality, computations are done only on finite sets. $x(t)$ is discretized, so is $\exp\left(i\frac{2\pi n}{p}t\right)$, and the limit of integration in (A.3) or summation in (A.4) is taken to be finite. Both become a discrete summation

$$F_n\left(x(t)\right) = F_n\left(x(k)\right) = \sum_{j=-k}^{k} x(j) \exp\left(i\frac{2\pi n}{p}j\right), \tag{A.8}$$

where k is a positive integer. Equation A.8 is known as the *discrete Fourier transform* (DFT), where $F_n(x(k))$ is usually replaced by X_n (Equation A.4 is also DFT, but (A.8) is a more appropriate form of DFT when implemented in electronic computers). For the signal segment or window $[-k, k]$, calculation of DFT (A.8) takes time on the order of k^2 ($O(k^2)$). A clever algorithm is capable of bringing down the time to $O(k \log k)$. This algorithm is known as *fast Fourier transform* (FFT). Nowadays, wherever DFT is implemented, it is done so only as FFT. By FFT on a signal segment (i.e., a digital signal of finite length), for each integer n we get a complex number X_n just as in the case of (A.3) or (A.4). Most signals of real life are not periodic and therefore are not ideal candidates for Fourier transform, in the sense that the Fourier series obtained by the FFT will not converge to the original signal at each time point. Yet FFT is calculated on many of them with results of great practical values. In all those cases, p is taken to be the length of the segment, that is, $p = 2k + 1$, with the assumption that this segment will keep repeating periodically to constitute the infinite-length signal (which is often not the case in reality). It is clear that there will be $2k + 1$ different X_n, as the values of X_n will keep repeating beyond $n = 2k + 1$. X_n and X_{-n} will be mirror images of each other. So, effectively, we are left with k different values of X_n. In other words, X_ns, the FFTs of the discrete signal $x(j)$, will give amplitude of the k different frequency components present in the signal.

Notice that in the above discussion after performing the Fourier transform (by FFT) on the time domain discrete signal of finite length $\{x(j)\}_{j=-k}^{k}$ or, for simplicity, $\{x(j)\}_{j=1}^{2k+1}$, the $2k + 1$ values of FFT that we get do not contain any information about the time of the signal (it has been summed as in (A.6) over time j to produce a result which does not contain j). They give only the amplitude of the frequency components as shown in (A.2) as they contain

only information on frequency. Through FFT, therefore, the signal is said to have transformed into frequency domain from the original time domain. However, we know the signal segment or window on which the FFTs were applied. The entire window in this case gives information on time, which is a large interval, and therefore this information is very blurred. We can make the time information sharper by taking a narrower window, but that will introduce more error into the FFT as the Fourier transformation is defined for $t = -\infty$ to $t = \infty$. This is known as *bias variance trade-off*. If we want to associate the frequencies with a precise time, the computation of amplitude associated with the frequencies will be hugely erroneous (biased in time). On the other hand, if we want the frequency information to be precise we must take a sufficiently long signal segment, which might result in too much blurring in time (variance in time).

Time-frequency analysis has predominant importance in EEG signal processing. For example, to analyze the cortical rhythms associated with different tasks performed by the brain, it is important to know what frequencies are dominating in the EEG signals coming from different parts of the head at the time of execution of certain tasks. Abnormality in brain rhythms may be indicative of pathological brain conditions like epilepsy and schizophrenia. One standard way of achieving it is through *short-time Fourier transformation* (STFT). In STFT, the FFT is performed in a short time window in order to keep the time location as sharp as possible. The FFT-based power spectrum is calculated which gives the intensity of each frequency band in which the source of the signal is oscillating. The short window is slid across the signal to carry out the measurement on the entire signal.

We have seen that Fourier transform converts a time domain signal into a frequency domain signal in the sense that we get $a_n + ib_n$, where a_n is the amplitude of the cosine wave with frequency n and b_n is the amplitude of the sine wave with frequency n (Equation A.2). Together they constitute the nth frequency component of the signal. So once we get the Fourier transformed output $a_n + ib_n$, we can readily construct the nth frequency component of the signal as $a_n \cos\left(\frac{2\pi n}{p} t\right) + b_n \sin\left(\frac{2\pi n}{p} t\right)$, provided the signal is periodic with period p and also satisfies Dirichlet's conditions as stated above. Notice that this component is in the time domain. Clearly, adding all such components together we get back the original signal by Equation A.2. All these have been encapsulated in one single operation, called *inverse Fourier transform* (iFT), as described below.

Assuming that the Fourier transform exists, that is, the right side of Equation A.3 converges to the left side for every value of t for the continuous case, we can write

$$x(t) = \int_{-p/2}^{p/2} F_n\left(x(\tau)\right) \exp\left(i \frac{2\pi t}{p} \tau\right) d\tau. \tag{A.9}$$

Equation A.9 is known as the iFT of the Fourier-transformed signal $F_n(x(t))$.

Note that here we have done all the calculations within $[-p/2, p/2]$. This is because of the assumption of the periodicity of the signal $x(t)$ with period p. This can be extended to nonperiodic signals, where the period tends to infinity (Proakis and Manolakis, 2007). In that case, the Fourier transform given by (A.3) will become

$$F_n\left(x\left(t\right)\right) = \lim_{p\to\infty} \int_{-p/2}^{p/2} x\left(t\right)\exp\left(-i\frac{2\pi n}{p}t\right)dt = \int_{-\infty}^{\infty} x\left(t\right)\exp\left(-i\frac{2\pi n}{p}t\right)dt. \qquad \text{(A.10)}$$

The iFT given by (A.9) becomes

$$x\left(t\right) = \lim_{p\to\infty} \int_{-p/2}^{p/2} F_n\left(x\left(\tau\right)\right)\exp\left(i\frac{2\pi t}{p}\tau\right)d\tau = \int_{-\infty}^{\infty} F_n\left(x\left(\tau\right)\right)\exp\left(i\frac{2\pi t}{p}\tau\right)d\tau. \qquad \text{(A.11)}$$

Although (A.10) and (A.11) are not valid from a purely mathematical point of view, from a computational point of view even the infinite period must be approximated by a (possibly large) finite period. In that sense, for all practical purposes, a nonperiodic signal can be taken as one with a very large period. If, in addition, the signal satisfies Dirichlet's conditions (which most real-life signals do), then the Fourier transform (A.10) and the iFT (A.11) are valid. In fact, this is the reason why FFT is performed on almost all types of signals in real life. Where discrete Fourier transform (given by (A.4), where $p \to \infty$ may be assumed for nonperiodic signals) takes $O(N^2)$ time to execute for a signal with N samples or time points, the FFT takes only $O(N\log_2 N)$ time to execute, which is a huge advantage for computational implementation. This is why Fourier transform has become synonymous to FFT.

Reference

Proakis, J. G. and D. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, 4th ed., Pearson Education Inc., New Delhi, India, 2007.

Appendix B: Wavelet Transformation

In Appendix A, we saw that by Fourier transformation a signal can be decomposed into sine and cosine series provided the signal satisfies certain conditions. This means the signal can be simulated as a linear superposition of signals from a large number of harmonic oscillators (such as tuning fork), each oscillating at a particular frequency but with different sizes (which is known as intensity, $\sqrt{a_n^2 + b_n^2}$ for the nth harmonic oscillator). Fourier transformations precisely determine this intensity or oscillator size. Fourier transformation can be carried out on any signal, but it will give appropriate results only when the signal is periodic and satisfies Dirichlet's conditions (Appendix A). Periodicity is, however, a very strict condition, which most real-life signals do not obey. EEG signals are particularly nonperiodic. Yet we need to decompose EEG signals to extract features from them. Fortunately, there are infinitely different ways signals can be decomposed into simpler, elementary waveforms. Sine and cosine waves are but two of them. The elementary waveforms in which a signal can be decomposed are called *wavelets*: $\sin(2\pi nt)$, $\cos(2\pi nt)$, $\exp(i2\pi nt)$, where $i = \sqrt{-1}$, are all examples of wavelets (wavelet transformations can be considered a generalization of Fourier transformations). But there are many other interesting and useful wavelets. Some of them are quite popular in EEG processing and in this appendix we will discuss them from a theoretical as well as an application point of view.

Here is an example of one of the simplest mathematical transformations (or transforms). Take any two-dimensional vector $\hat{\mathbf{a}} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}$, where x and y are scalar quantities, and $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are unit vectors along X and Y directions, respectively. Consider the situation where we have an arbitrary two-dimensional vector $\hat{\mathbf{a}}$ and we do not know x and y; how do we determine x and y? We find them by $x = \hat{\mathbf{a}} \cdot \hat{\mathbf{i}}$ and $y = \hat{\mathbf{a}} \cdot \hat{\mathbf{j}}$. Here the dot product or the inner product is a transformation operation, which helps to find the scalars involved in describing the vector $\hat{\mathbf{a}}$ and therefore completely determines $\hat{\mathbf{a}}$. Both Fourier and wavelet transformations are applied on a signal in order to extract simpler components from that signal, so that the entire signal can be described in terms of those simpler components. The dot product or the inner product is generalized by an integration operation that we have already seen for Fourier transformation in Appendix A. In case of $\hat{\mathbf{a}}$ by the dot products or for a periodic signal satisfying Dirichlet's conditions (Appendix A) by the Fourier transforms, the vector or the signal, respectively, can be obtained completely. If the signal is not periodic or does not satisfy Dirichlet's conditions (or certain other conditions in place of Dirichlet's conditions), the signal cannot be recovered from its Fourier-transformed values. But the Fourier transform may still give us some useful information about the signal, for

example, estimation of power spectrum. Better approximation of the signal is possible in these cases by appropriate wavelet transformations.

Fourier transforms are only one class of transforms, namely sinusoidal transforms. Wavelet transform is, on the other hand, the generic name for an infinite class of transformations. An appropriate class is to be chosen for the signal(s) at hand depending on which type of information is to be retrieved from the signal(s). Unfortunately, there is no mathematical theory for this choice. It depends entirely on the practitioner's wisdom. The key to wavelet transformation is to capture the frequencies present in a nonperiodic signal at a given time interval (for a periodic signal all the frequency components of the signal remain present for the entire duration of the signal).

One way of doing this is to take the short-time Fourier transform (STFT). Whenever we need to know the frequency components of the signal at time t, we have to take an interval $[t-\delta, t+\delta]$ and do the Fourier transform on that interval. However, the Fourier transform is defined on an interval $(-\infty, \infty)$ and therefore the shorter the δ is, the more erroneous the STFT will be. To circumvent this problem, continuous wavelet transform, as described here, was introduced:

$$\Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{s}} \int x(t) \psi^* \left(\frac{t-\tau}{s} \right) dt, \tag{B.1}$$

where $\Psi_x^{\psi}(\tau, s)$ is called *wavelet coefficient* with respect to the wavelet transformation by the *wavelet function* ψ of the signal $x(t) \cdot \psi^*$ is the complex conjugate of $\psi \cdot \tau$ and s are parameters whose significance will soon be made clear. In case of Fourier transformation, ψ can be only of the form $\psi(n,t) = \exp(j2\pi nt)$, where $j = \sqrt{-1}$ (see Appendix A for details). But in case of wavelet transformation, ψ can take infinitely different forms. Depending on the form, it can have different names, such as Harr wavelet or Morlet wavelet. In fact, ψ is called the *wavelet* or small wave with which the signal $x(t)$ is to be locally approximated. This approximation is made in terms of 'at what time instance the signal is having which frequency component'. This is in short called *time-frequency analysis*. Of course, a time domain signal will usually have multiple frequencies at each time instance. The analysis is to be run as many times as there are frequencies in a time instance so that we can get the value of the wavelet coefficient associated with each of the frequencies involved. This is explained in Figure B.1.

In case of STFT, the Fourier transform is calculated for each frequency between, say, time 0 and time 10, and values are recorded in all the boxes in the first column of Figure B.1. The first column will contain the value for a frequency of 1 Hz, the second for 2 Hz, and so on. Similarly, calculations can be made for time 10 to time 20, for time 20 to time 30, and so on, to eventually fill up all the boxes in the figure. In Figure B.1, we have only accommodated frequencies up to 5 Hz, but the plot can be extended in a similar way to

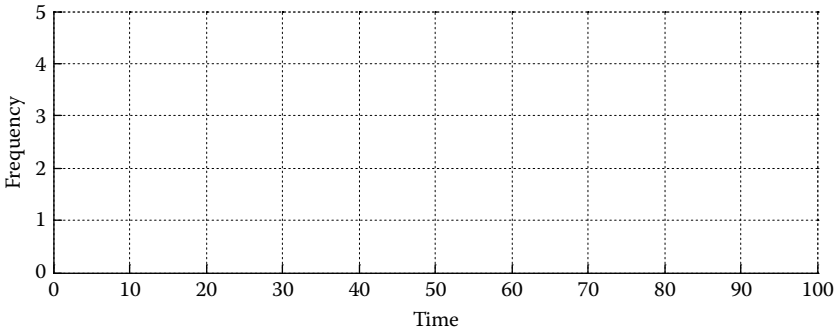


FIGURE B.1
Two-dimensional time-frequency analysis framework for a one-dimensional time domain signal. Frequency of the signal is to be determined at each time interval and the value is to be put in the appropriate box.

accommodate frequencies up to any range. A zero value in a box will indicate that frequency in that particular time interval is absent from that signal. Note that we are determining frequency over a time interval, not on a particular time instance or time point, because that is impossible to determine in the same way as momentum and position of an atomic or subatomic particle is impossible to determine perfectly at the same time in physics, known as Heisenborg's *uncertainty principle*. This happens because Fourier transform is to be calculated over a time interval; it cannot be calculated on a single time point.

The usual EEG signals will be so complicated that the box size of Figure B.1 will keep changing over time. This happens because the frequencies are changing dynamically over a given interval, which is a typical characteristic of a nonperiodic signal. To capture this phenomenon, STFT is not an appropriate tool. We should go back to Equation B.1 and have a closer look at it: $\psi^*\left(\frac{t-\tau}{s}\right)$ is a function which is defined on the time axis (i.e., time-dependent function) centered at $t=\tau$, and scaled by a factor $1/s$. That is, the distance on the time axis $t-\tau$ can be compressed (if $\frac{1}{s}<1$) or stretched (if $\frac{1}{s}>1$) before being mapped under ψ^*. Both τ and s are continuous parameters and called *shift* and *scaling* parameters, respectively. This is why ψ is called continuous wavelet. However, at the time of implementation by an electronic computer, both must be discretized. The smaller the step size, the more accurate will be the computation, but with a greater computational cost.

First, a τ has to be fixed and then the wavelet coefficients for that τ for different values of s are to be calculated for signal $x(t)$ (usually within a large enough window with midpoint τ) according to Equation B.1. As s becomes large, $1/s$ becomes small, and $t-\tau$ is contracted. Therefore, under the mapping

by ψ^*, it becomes amplified or dilated. Similarly, when s becomes small, $t-\tau$ is expanded and under the mapping by ψ^*, it becomes compressed or contracted. Now, if, where (in the time interval) the frequency of the signal is high, that portion is convoluted with ψ^* for a small s (i.e., with a contraction wavelet), the resulting numerical value of the wavelet coefficient (according to Equation B.1) will normally be high (in a statistical sense, rather than in a pure mathematical sense). Similarly, if, where in a time interval the frequency of the signal is low, that portion is convoluted with ψ^* for a large s (i.e., with a dilation wavelet), the resulting numerical value of the wavelet coefficient will normally be high. In other words, the scaling adjusts the wavelet according to the local frequency of the time domain signal. The original wavelet of the form $\psi\left(\frac{t-\tau}{s}\right)$ is called *mother wavelet* and any scaled (for a value of s) or translated (for a value of τ) (generic) mother wavelet becomes a *daughter wavelet*. Like in Fourier transform, where the signal is to be convoluted with all the Fourier components (sine and cosine for all frequencies), in wavelet transform the signal is to be convoluted with all the daughter wavelets, once a particular mother wavelet is chosen. We have presented one instance of continuous wavelet transformation in Section 5.1.1.2, where we have described a wavelet transformation–based phase synchronization measure between two time domain EEG signals. Here the mother wavelet was a Morlet (or Gabor) wavelet: $\psi^*\left(\frac{t-\tau}{s}\right)$ and scale was the inverse frequency, that is, $s=1/f$.

Discrete wavelet transform is one in which the mother wavelet function ψ comes with τ and s that are allowed to take integer values only (in discretized continuous wavelet transformation τ and s can take on any value). Notice that a signal's overall (or global) shape depends more on the low-frequency components and the local details depend more on the high-frequency components. So, the idea is to divide the signal into high-frequency and low-frequency components and then do the same for the low-frequency component to again get one high-frequency and one low-frequency component and continue to do it with the low-frequency component at each stage (this has already been shown in Section 6.1.2). This is defined for a digital signal only and therefore takes the sampling frequency of the signal to start with. If it is n, then apply a low-pass and a high-pass filter with cutoff frequency $\frac{n}{2}$ in each case and carry on iteratively for the low-pass-filtered signal obtained at each stage.

Index

C

D

E

M

N

T

U

V

W

Z