The Digital Signal Processing Handbook

SECOND EDITION

Wireless, Networking, Radar, Sensor Array Processing, and Nonlinear Signal Processing

EDITOR-IN-CHIEF
Vijay K. Madisetti

CRC Press
Taylor & Francis Group

The Digital Signal Processing Handbook

SECOND EDITION

Wireless, Networking, Radar, Sensor Array Processing, and Nonlinear Signal Processing

EDITOR-IN-CHIEF

Vijay K. Madisetti

CRC Press
Taylor & Francis Group
Boca Raton London New York

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

The Electrical Engineering Handbook Series

Series Editor

Richard C. Dorf
University of California, Davis

Titles Included in the Series

The Handbook of Ad Hoc Wireless Networks, Mohammad Ilyas
The Avionics Handbook, Second Edition, Cary R. Spitzer
The Biomedical Engineering Handbook, Third Edition, Joseph D. Bronzino
The Circuits and Filters Handbook, Second Edition, Wai-Kai Chen
The Communications Handbook, Second Edition, Jerry Gibson
The Computer Engineering Handbook, Vojin G. Oklobdzija
The Control Handbook, William S. Levine
The CRC Handbook of Engineering Tables, Richard C. Dorf
The Digital Avionics Handbook, Second Edition Cary R. Spitzer
The Digital Signal Processing Handbook, Second Edition, Vijay K. Madisetti
The Electrical Engineering Handbook, Second Edition, Richard C. Dorf
The Electric Power Engineering Handbook, Second Edition, Leonard L. Grigsby
The Electronics Handbook, Second Edition, Jerry C. Whitaker
The Engineering Handbook, Third Edition, Richard C. Dorf
The Handbook of Formulas and Tables for Signal Processing, Alexander D. Poularikas
The Handbook of Nanoscience, Engineering, and Technology, Second Edition
William A. Goddard, III, Donald W. Brenner, Sergey E. Lyshevski, and Gerald J. Iafrate
The Handbook of Optical Communication Networks, Mohammad Ilyas and
Hussein T. Mouftah
The Industrial Electronics Handbook, J. David Irwin
The Measurement, Instrumentation, and Sensors Handbook, John G. Webster
The Mechanical Systems Design Handbook, Osita D.I. Nwokah and Yidirim Hurmuzlu
The Mechatronics Handbook, Second Edition, Robert H. Bishop
The Mobile Communications Handbook, Second Edition, Jerry D. Gibson
The Ocean Engineering Handbook, Ferial El-Hawary
The RF and Microwave Handbook, Second Edition, Mike Golio
The Technology Management Handbook, Richard C. Dorf
The Transforms and Applications Handbook, Second Edition, Alexander D. Poularikas
The VLSI Handbook, Second Edition, Wai-Kai Chen

The Digital Signal Processing Handbook, Second Edition

Digital Signal Processing Fundamentals
Video, Speech, and Audio Signal Processing and Associated Standards
Wireless, Networking, Radar, Sensor Array Processing, and Nonlinear Signal Processing

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

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-4604-5 (Hardback)

Library of Congress Cataloging-in-Publication Data

Wireless, networking, radar, sensor array processing, and nonlinear signal processing / Vijay K. Madisetti.
p. cm.
"Second edition of the DSP Handbook has been divided into three parts."
Includes bibliographical references and index.
ISBN 978-1-4200-4604-5 (alk. paper)
1. Signal processing--Digital techniques. 2. Wireless communication systems. 3. Array processors. 4. Computer networks. 5. Radar. I. Madisetti, V. (Vijay) II. Digital signal processing handbook. III. Title.

TK5102.9.W555 2009
621.382'2--dc22 2009022597

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

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

Contents

Preface

Digital signal processing (DSP) is concerned with the theoretical and practical aspects of representing information-bearing signals in a digital form and with using computers, special-purpose hardware and software, or similar platforms to extract information, process it, or transform it in useful ways. Areas where DSP has made a significant impact include telecommunications, wireless and mobile communications, multimedia applications, user interfaces, medical technology, digital entertainment, radar and sonar, seismic signal processing, and remote sensing, to name just a few.

Given the widespread use of DSP, a need developed for an authoritative reference, written by the top experts in the world, that would provide information on both theoretical and practical aspects in a manner that was suitable for a broad audience—ranging from professionals in electrical engineering, computer science, and related engineering and scientific professions to managers involved in technical marketing, and to graduate students and scholars in the field. Given the abundance of basic and introductory texts on DSP, it was important to focus on topics that were useful to engineers and scholars without overemphasizing those topics that were already widely accessible. In short, the DSP handbook was created to be relevant to the needs of the engineering community.

A task of this magnitude could only be possible through the cooperation of some of the foremost DSP researchers and practitioners. That collaboration, over 10 years ago, produced the first edition of the successful DSP handbook that contained a comprehensive range of DSP topics presented with a clarity of vision and a depth of coverage to inform, educate, and guide the reader. Indeed, many of the chapters, written by leaders in their field, have guided readers through a unique vision and perception garnered by the authors through years of experience.

The second edition of the DSP handbook consists of *Digital Signal Processing Fundamentals*; *Video, Speech, and Audio Signal Processing and Associated Standards*; and *Wireless, Networking, Radar, Sensor Array Processing, and Nonlinear Signal Processing* to ensure that each part is dealt with in adequate detail, and that each part is then able to develop its own individual identity and role in terms of its educational mission and audience. I expect each part to be frequently updated with chapters that reflect the changes and new developments in the technology and in the field. The distribution model for the DSP handbook also reflects the increasing need by professionals to access content in electronic form anywhere and at anytime.

Wireless, Networking, Radar, Sensor Array Processing, and Nonlinear Signal Processing, as the name implies, provides a comprehensive coverage of the foundations of signal processing related to wireless, radar, space–time coding, and mobile communications, together with associated applications to networking, storage, and communications.

This book needs to be continuously updated to include newer aspects of these technologies, and I look forward to suggestions on how this handbook can be improved to serve you better.

MATLAB® is a registered trademark of The MathWorks, Inc. For product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01760-2098 USA
Tel: 508 647 7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

Editor

Vijay K. Madisetti is a professor in the School of Electrical and Computer Engineering at the Georgia Institute of Technology in Atlanta. He teaches graduate and undergraduate courses in digital signal processing and computer engineering, and leads a strong research program in digital signal processing, telecommunications, and computer engineering.

Dr. Madisetti received his BTech (Hons) in electronics and electrical communications engineering in 1984 from the Indian Institute of Technology, Kharagpur, India, and his PhD in electrical engineering and computer sciences in 1989 from the University of California at Berkeley.

He has authored or edited several books in the areas of digital signal processing, computer engineering, and software systems, and has served extensively as a consultant to industry and the government. He is a fellow of the IEEE and received the 2006 Frederick Emmons Terman Medal from the American Society of Engineering Education for his contributions to electrical engineering.

Contributors

Naofal Al-Dhahir
Department of Electrical Engineering
The University of Texas at Dallas
Richardson, Texas

Mishal Al-Gharabally
Electrical Engineering Department
College of Engineering and Petroleum
Safat, Kuwait

K. Balemarthy
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Victor A. N. Barroso
Department of Electrical and Computer Engineering
Instituo Superior Tecnico
Instituto de Sistemas e Robótica
Lisbon, Portugal

John R. Barry
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Robert J. Baxley
Georgia Tech Research Institute
Atlanta, Georgia

Kevin M. Buckley
Department of Electrical and Computer Engineering
Villanova University
Villanova, Pennsylvania

Robert Calderbank
Department of Electrical Engineering
Princeton University
Princeton, New Jersey

Chaitali Chakrabarti
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, Arizona

Jean-Yves Chouinard
Department of Electronic Engineering and Computer Science
Laval University
Quebec, Quebec, Canada

Jimmy Chui
Department of Electrical Engineering
Princeton University
Princeton, New Jersey

Kevin M. Cuomo
Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, Massachusetts

Pankaj Das
Department of Electrical and Computer Engineering
University of California
San Diego, California

Sushanta Das
Phillips Research N.A.
New York, New York

J. Debardelaben
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

R. D. DeGroat
Broadcom Corporation
Denver, Colorado

N. Desai
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Suhas Diggavi
Ecole Polytechnique
Lausanne, Switzerland

E. M. Dowling
Department of Electrical Engineering
The University of Texas at Dallas
Richardson, Texas

Lan-Rong Dung
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

T. Egolf
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Atef Z. Elsherbeni
Department of Electrical Engineering
University of Mississippi
Oxford, Mississippi

S. Famorzadeh
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Daniel R. Fuhrmann
Department of Electrical and System Engineering
Washington University
St. Louis, Missouri

Egemen Gönen
Globalstar
San Jose, California

Martin Haardt
Communication Research Laboratory
Ilmenau University of Technology
Ilmenau, Germany

Emile A. Hendriks
Information and Communication Theory Group
Delft University of Technology
Delft, the Netherlands

R. Hezar
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Steven H. Isabelle
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

Yong-kyu Jung
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

D. I. Kaklamani
Department of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece

D. A. Kateros
Department of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece

Mostafa Kaveh
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, Minnesota

A. Kavipurapu
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Brian Kelley
Department of Electrical and Computer Engineering
The University of Texas at San Antonio
San Antonio, Texas

M. Khan
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Yuan Lin
Advanced Computer Architecture Laboratory
University of Michigan at Ann Arbor
Ann Arbor, Michigan

D. A. Linebarger
Department of Electrical Engineering
The University of Texas at Dallas
Richardson, Texas

K. J. Ray Liu
Department of Electrical and Computer Engineering
University of Maryland
College Park, Maryland

Xiaoli Ma
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Vijay K. Madisetti
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Bassem R. Mahafza
Deceibel Research, Inc.
Huntsville, Alabama

Scott Mahlke
Advanced Computer Architecture Laboratory
University of Michigan at Ann Arbor
Ann Arbor, Michigan

Petros Maragos
Department of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece

Cherian P. Mathews
Department of Electrical and Computer Engineering
University of the Pacific
Stockton, California

Jerry M. Mendel
Department of Electrical Engineering
University of Southern California
Los Angeles, California

David L. Milliner
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

José M. F. Moura
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, Pennsylvania

Trevor Mudge
Advanced Computer Architecture Laboratory
University of Michigan at Ann Arbor
Ann Arbor, Michigan

Arye Nehorai
Department of Electrical and Computer Engineering
The University of Illinois at Chicago
Chicago, Illinois

Alan V. Oppenheim
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

Eytan Paldi
Department of Mathematics
Israel Institute of Technology
Technion City, Haifa, Israel

C. B. Papadias
Broadband Wireless
Athens Information Technology
Peania Attikis, Greece

Panos Papamichalis
Texas Instruments
Dallas, Texas

A. Paulraj
Department of Electrical Engineering
Stanford University
Stanford, California

Athina P. Petropulu
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, Pennsylvania

M. Pettigrew
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

G. N. Prezerakos
Department of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece

and

Technological Education Institute of Piraeus
Athens, Greece

Javier Ramos
Department of Signal Processing and Communications
Universidad Rey Juan Carlos
Madrid, Spain

André Redert
Philips Research Europe
Eindhoven, the Netherlands

Zoltan Safar
Department of Innovation
IT University of Copenhagen
Copenhagen, Denmark

Sangwon Seo
Advanced Computer Architecture Laboratory
University of Michigan at Ann Arbor
Ann Arbor, Michigan

Andrew C. Singer
Sanders (A Lockhead Martin Company)
Manchester, New Hampshire

Mohanned O. Sinnokrot
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Brian J. Smith
U.S. Army Aviation and Missile Command
Redstone Arsenal, Alabama

P. Stoica
Information Technology Department
Uppsala University
Uppsala, Sweden

Weifeng Su
Department of Electrical Engineering
State University of New York at Buffalo
Buffalo, New York

Barry Van Veen
Department of Electrical and Computer Engineering
University of Wisconsin
Madison, Wisconsin

Mats Viberg
Department of Signal and Systems
Chalmers University of Technology
Goteborg, Sweden

Hong Wang
Department of Electrical and Computer Engineering
Syracuse University
Syracuse, New York

Xianbin Wang
Department of Electrical and Computer Engineering
University of Western Ontario
London, Ontario, Canada

Douglas B. Williams
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Mark Woh
Advanced Computer Architecture Laboratory
University of Michigan at Ann Arbor
Ann Arbor, Michigan

M. Wong
Department of Electrical and Computer Engineering
McMaster University
Hamilton, Ontario, Canada

Gregory W. Wornell
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

Q. Wu
CELWAVE
Claremont, North Carolina

Yiyan Wu
Communications Research Centre
Ottawa, Ontario, Canada

D. A. Zarbouti
Department of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece

Wei Zhang
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

G. Tong Zhou
Department of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

Michael D. Zoltowski
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana

I

Sensor Array Processing

Mostafa Kaveh
University of Minnesota

A SENSOR ARRAY SYSTEM CONSISTS OF a number of spatially distributed elements, such as dipoles, hydrophones, geophones or microphones, followed by receivers and a processor. The array samples propagate wavefields in time and space. The receivers and the processor vary in mode of implementation and complexity according to the types of signals encountered, the desired operation, and the adaptability of the array. For example, the array may be narrowband or wideband and the processor may be for determining the directions of the sources of signals or for beamforming to reject interfering signals and to enhance the quality of the desired signal in a communication system. The broad range of applications and the multifaceted nature of technical challenges for modern array signal processing have provided a fertile ground for contributions by and collaborations among researchers and practitioners from many disciplines, particularly those from the signal processing, statistics, and numerical linear algebra communities.

The following chapters present a sampling of the latest theory, algorithms, and applications related to array signal processing. The range of topics and algorithms include some which have been in use for more than a decade as well as some which are results of active current research. The sections on applications give examples of current areas of significant research and development.

Modern array signal processing often requires the use of the formalism of complex variables in modeling received signals and noise. Chapter 1 provides an introduction to complex random processes which are useful for bandpass communication systems and arrays. A classical use for arrays of sensors is to exploit the differences in the location (direction) of sources of transmitted signals to perform spatial filtering. Such techniques are reviewed in Chapter 2.

Another common use of arrays is the estimation of informative parameters about the wavefields impinging on the sensors. The most common parameter of interest is the direction of arrival (DOA) of a wave. Subspace techniques have been advanced as a means of estimating the DOAs of sources, which are very close to each other, with high accuracy. The large number of developments in such techniques is reflected in the topics covered in Chapters 3 through 7. Chapter 3 gives a general overview of subspace processing for direction finding, while Chapter 4 discusses a particular type of subspace algorithm that is extended to sensing of azimuth and elevation angles with planar arrays. Most estimators assume

knowledge of the needed statistical characteristics of the measurement noise. This requirement is relaxed in the approach given in Chapter 5. Chapter 6 extends the capabilities of traditional sensors to those which can measure the complete electric and magnetic field components and provides estimators which exploit such information. When signal sources move, or when computational requirements for real-time processing prohibit batch estimation of the subspaces, computationally efficient adaptive subspace updating techniques are called for. Chapter 7 presents many of the recent techniques that have been developed for this purpose. Before subspace methods are used for estimating the parameters of the waves received by an array, it is necessary to determine the number of sources which generate the waves. This aspect of the problem, often termed detection, is discussed in Chapter 8.

An important area of application for arrays is in the field of communications, particularly as it pertains to emerging mobile and cellular systems. Chapter 9 gives an overview of a number of techniques for improving the reception of signals in mobile systems, while Chapter 10 considers problems that arise in beamforming in the presence of multipath signals—a common occurrence in mobile communications. Chapter 12 discusses radar systems that employ sensor arrays, thereby providing the opportunity for space–time signal processing for improved resolution and target detection.

1

Complex Random Variables and Stochastic Processes

Daniel R. Fuhrmann
Washington University

1.1 Introduction

Much of modern digital signal processing is concerned with the extraction of information from signals which are noisy, or which behave randomly while still revealing some attribute or parameter of a system or environment under observation. The term in popular use now for this kind of computation is "statistical signal processing," and much of this handbook is devoted to this very subject. Statistical signal processing is classical statistical inference applied to problems of interest to electrical engineers, with the added twist that answers are often required in "real time," perhaps seconds or less. Thus, computational algorithms are often studied hand-in-hand with statistics.

One thing that separates the phenomena electrical engineers study from that of agronomists, economists, or biologists, is that the data they process are very often complex; that is, the data points come in pairs of the form $x + jy$, where x is called the real part, y the imaginary part, and $j = \sqrt{-1}$. Complex numbers are entirely a human intellectual creation: there are no complex physical measurable quantities such as time, voltage, current, money, employment, crop yield, drug efficacy, or anything else. However, it is possible to attribute to physical phenomena an underlying mathematical model that associates complex causes with real results. Paradoxically, the introduction of a complex-number-based theory can often simplify mathematical models.

Beyond their use in the development of analytical models, complex numbers often appear as actual data in some information processing systems. For representation and computation purposes, a complex number is nothing more than an ordered pair of real numbers. One just mentally attaches the "j" to one of the two numbers, then carries out the arithmetic or signal processing that this interpretation of the data implies.

One of the most well-known systems in electrical engineering that generates complex data from real measurements is the quadrature, or IQ, demodulator, shown in Figure 1.1. The theory behind this system is as follows. A real bandpass signal, with bandwidth small compared to its center frequency, has the form

$$s(t) = A(t)\cos(\omega_c t + \phi(t)), \tag{1.1}$$

where

ω_c is the center frequency

$A(t)$ and $\phi(t)$ are the amplitude and angle modulation, respectively

By viewing $A(t)$ and $\phi(t)$ together as the polar coordinates for a complex function $g(t)$, i.e.,

$$g(t) = A(t)e^{j\phi(t)}, \tag{1.2}$$

we imagine that there is an underlying "complex modulation" driving the generation of $s(t)$, and thus

$$s(t) = \mathrm{Re}\{g(t)e^{j\omega_c t}\}. \tag{1.3}$$

Again, $s(t)$ is physically measurable, while $g(t)$ is a mathematical creation. However, the introduction of $g(t)$ does much to simplify and unify the theory of bandpass communication. It is often the case that information to be transmitted via an electronic communication channel can be mapped directly into the magnitude and phase, or the real and imaginary parts, of $g(t)$. Likewise, it is possible to demodulate $s(t)$, and thus "retrieve" the complex function $g(t)$ and the information it represents. This is the purpose of the quadrature demodulator shown in Figure 1.1. In Section 1.2, we will examine in some detail the operation of this demodulator, but for now note that it has one real input and two real outputs, which are interpreted as the real and imaginary parts of an information-bearing complex signal.

Any application of statistical inference requires the development of a probabilistic model for the received or measured data. This means that we imagine the data to be a "realization" of a multivariate random variable, or a stochastic process, which is governed by some underlying probability space of which we have incomplete knowledge. Thus, the purpose of this section is to give an introduction to probabilistic models for complex data. The topics covered are second-order stochastic processes and their

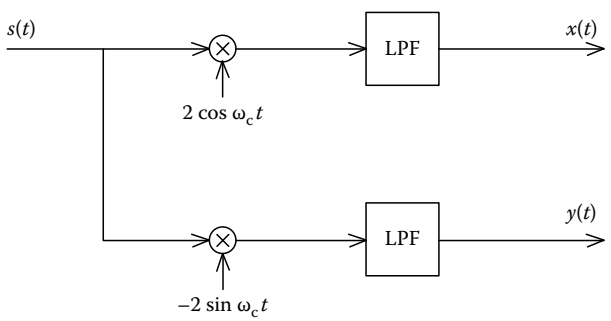

FIGURE 1.1 Quadrature demodulator.

complex representations, the multivariate complex Gaussian distribution, and related distributions which appear in statistical tests. Special attention will be paid to a particular class of random variables, called "circular" complex random variables. Circularity is a type of symmetry in the distributions of the real and imaginary parts of complex random variables and stochastic processes, which can be physically motivated in many applications and is almost always assumed in the statistical signal processing literature. Complex representations for signals and the assumption of circularity are particularly useful in the processing of data or signals from an array of sensors, such as radar antennas. The reader will find them used throughout this chapter of the handbook.

1.2 Complex Envelope Representations of Real Bandpass Stochastic Processes

1.2.1 Representations of Deterministic Signals

The motivation for using complex numbers to represent real phenomena, such as radar or communication signals, may be best understood by first considering the complex envelope of a real deterministic finite-energy signal.

Let $s(t)$ be a real signal with a well-defined Fourier transform $S(\omega)$. We say that $s(t)$ is bandlimited if the support of $S(\omega)$ is finite, that is,

$$S(\omega)\begin{cases} = 0 & \omega \notin B \\ \neq 0 & \omega \notin B \end{cases}, \tag{1.4}$$

where B is the frequency band of the signal, usually a finite union of intervals on the ω-axis such as

$$B = [-\omega_2, -\omega_1] \cup [\omega_1, \omega_2]. \tag{1.5}$$

The Fourier transform of such a signal is illustrated in Figure 1.2.

Since $s(t)$ is real, the Fourier transform $S(\omega)$ exhibits conjugate symmetry, i.e., $S(-\omega) = S^*(\omega)$. This implies that knowledge of $S(\omega)$, for $\omega \geq 0$ only, is sufficient to uniquely identify $s(t)$.

The complex envelope of $s(t)$, which we denote $g(t)$, is a frequency-shifted version of the complex signal whose Fourier transform is $S(\omega)$ for positive ω, and 0 for negative ω. It is found by the operation indicated graphically by the diagram in Figure 1.3, which could be written

$$g(t) = \mathrm{LPF}\{2s(t)e^{-j\omega_c t}\}. \tag{1.6}$$

ω_c is the center frequency of the band B

"LPF" represents an ideal lowpass filter whose bandwidth is greater than half the bandwidth of $s(t)$, but much less than $2\omega_c$

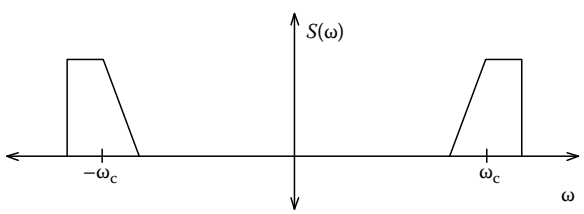

FIGURE 1.2 Fourier transform of a bandpass signal.

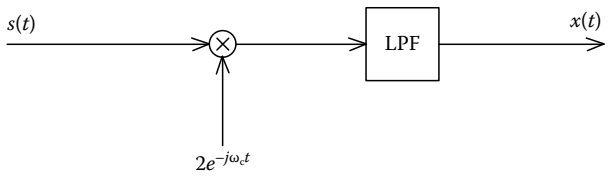

FIGURE 1.3 Quadrature demodulator.

The Fourier transform of $g(t)$ is given by

$$G(\omega) = \begin{cases} 2S(\omega - \omega_c) & |\omega| < BW \\ 0 & \text{otherwise} \end{cases}. \tag{1.7}$$

The Fourier transform of $g(t)$, for $s(t)$ as given in Figure 1.2, is shown in Figure 1.4.

The inverse operation which gives $s(t)$ from $g(t)$ is

$$s(t) = \text{Re}\{g(t)e^{j\omega_c t}\}. \tag{1.8}$$

Our interest in $g(t)$ stems from the information it represents. Real bandpass processes can be written in the form

$$s(t) = A(t)\cos(\omega_c t + \phi(t)), \tag{1.9}$$

where $A(t)$ and $\phi(t)$ are slowly varying functions relative to the unmodulated carrier $\cos(\omega_c t)$, and carry information about the signal source. From the complex envelope representation (Equation 1.3), we know that

$$g(t) = A(t)e^{j\phi(t)} \tag{1.10}$$

and hence $g(t)$, in its polar form, is a direct representation of the information-bearing part of the signal.

In what follows we will outline a basic theory of complex representations for real stochastic processes, instead of the deterministic signals discussed above. We will consider representations of second-order stochastic processes, those with finite variances and correlations and well-defined spectral properties. Two classes of signals will be treated separately: those with finite energy (such as radar signals) and those with finite power (such as radio communication signals).

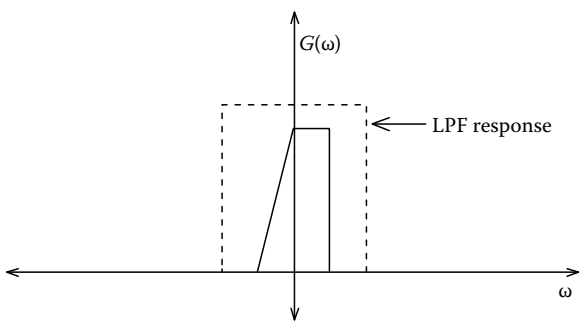

FIGURE 1.4 Fourier transform of the complex representation.

1.2.2 Finite-Energy Second-Order Stochastic Processes

Let $\mathbf{x}(t)$ be a real, second-order stochastic process, with the defining property

$$E\{\mathbf{x}^2(t)\} < \infty, \quad \text{all } t. \tag{1.11}$$

Furthermore, let $\mathbf{x}(t)$ be finite-energy, by which we mean

$$\int_{-\infty}^{\infty} E\{\mathbf{x}^2(t)\}\mathrm{d}t < \infty. \tag{1.12}$$

The autocorrelation function for $\mathbf{x}(t)$ is defined as

$$R_{\mathrm{xx}}(t_1, t_2) = E\{\mathbf{x}(t_1)\mathbf{x}(t_2)\}, \tag{1.13}$$

and from Equation 1.11 and the Cauchy–Schwartz inequality we know that R_{xx} is finite for all t_1, t_2.

The bifrequency energy spectral density function is

$$S_{\mathrm{xx}}(\omega_1, \omega_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_{\mathrm{xx}}(t_1, t_2)\mathrm{e}^{-j\omega_1 t_1}\mathrm{e}^{+j\omega_2 t_2}\mathrm{d}t_1\mathrm{d}t_2. \tag{1.14}$$

It is assumed that $S_{\mathrm{xx}}(\omega_1, \omega_2)$ exists and is well defined. In an advanced treatment of stochastic processes (e.g., Loeve [1]) it can be shown that $S_{\mathrm{xx}}(\omega_1, \omega_2)$ exists if and only if the Fourier transform of $\mathbf{x}(t)$ exists with probability 1; in this case, the process is said to be "harmonizable."

If $\mathbf{x}(t)$ is the input to a linear time-invariant (LTI) system $\mathbf{H}$, and $\mathbf{y}(t)$ is the output process, as shown in Figure 1.5, then $\mathbf{y}(t)$ is also a second-order finite-energy stochastic process. The bifrequency energy spectral density of $\mathbf{y}(t)$ is

$$S_{\mathrm{yy}}(\omega_1, \omega_2) = H(\omega_1)H^*(\omega_2)S_{\mathrm{xx}}(\omega_1, \omega_2). \tag{1.15}$$

This last result aids in a natural interpretation of the function $S_{\mathrm{xx}}(\omega, \omega)$, which we denote as the "energy spectral density." For any process, the total energy E_{x} is given by

$$E_{\mathrm{x}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{\mathrm{xx}}(\omega, \omega)\mathrm{d}\omega. \tag{1.16}$$

If we pass $\mathbf{x}(t)$ through an ideal filter whose frequency response is 1 in the band B and 0 elsewhere, then the total energy in the output process is

$$E_{\mathrm{y}} = \frac{1}{2\pi}\int_{B} S_{\mathrm{xx}}(\omega, \omega)\mathrm{d}\omega. \tag{1.17}$$

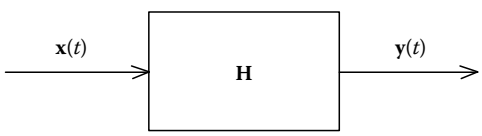

FIGURE 1.5 LTI system with stochastic input and output.

This says that the energy in the stochastic process $\mathbf{x}(t)$ can be partitioned into different frequency bands, and the energy in each band is found by integrating $S_{xx}(\omega, \omega)$ over the band.

We can define a "bandpass" stochastic process, with band B, as one that passes undistorted through an ideal filter $\mathbf{H}$ whose frequency response is 1 within the frequency band and 0 elsewhere. More precisely, if $\mathbf{x}(t)$ is the input to an ideal filter $\mathbf{H}$, and the output process $\mathbf{y}(t)$ is equivalent to $\mathbf{x}(t)$ in the mean-square sense, that is,

$$E\{(\mathbf{x}(t) - \mathbf{y}(t))^2\} = 0, \quad \text{all } t, \tag{1.18}$$

then we say that $\mathbf{x}(t)$ is a bandpass process with frequency band equal to the passband of $\mathbf{H}$. This is equivalent to saying that the integral of $S_{xx}(\omega_1, \omega_2)$ outside of the region $\omega_1, \omega_2 \in B$ is 0.

1.2.3 Second-Order Complex Stochastic Processes

A "complex" stochastic process $\mathbf{z}(t)$ is one given by

$$\mathbf{z}(t) = \mathbf{x}(t) + j\mathbf{y}(t) \tag{1.19}$$

where the real and imaginary parts, $\mathbf{x}(t)$ and $\mathbf{y}(t)$, respectively, are any two stochastic processes defined on a common probability space. A finite-energy, second-order complex stochastic process is one in which $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are both finite-energy, second-order processes, and thus have all the properties given above. Furthermore, because the two processes have a joint distribution, we can define the "cross-correlation function"

$$R_{xy}(t_1, t_2) = E\{\mathbf{x}(t_1)\mathbf{y}(t_2)\}. \tag{1.20}$$

By far the most widely used class of second-order complex processes in signal processing is the class of "circular" complex processes. A circular complex stochastic process is one with the following two defining properties:

$$R_{xx}(t_1, t_2) = R_{yy}(t_1, t_2) \tag{1.21}$$

and

$$R_{xy}(t_1, t_2) = -R_{yx}(t_1, t_2), \quad \text{all } t_1, t_2. \tag{1.22}$$

From Equations 1.21 and 1.22 we have that

$$E\{\mathbf{z}(t_1)\mathbf{z}^*(t_2)\} = 2R_{xx}(t_1, t_2) + 2jR_{yx}(t_1, t_2) \tag{1.23}$$

and furthermore

$$E\{\mathbf{z}(t_1)\mathbf{z}(t_2)\} = 0, \quad \text{all } t_1, t_2. \tag{1.24}$$

This implies that all of the joint second-order statistics for the complex process $\mathbf{z}(t)$ are represented in the function

$$R_{zz}(t_1, t_2) = E\{\mathbf{z}(t_1)\mathbf{z}^*(t_2)\} \tag{1.25}$$

which we define unambiguously as the autocorrelation function for $\mathbf{z}(t)$. Likewise, the bifrequency spectral density function for $\mathbf{z}(t)$ is given by

$$S_{zz}(\omega_1, \omega_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_{zz}(t_1, t_2) e^{-j\omega_1 t_1} e^{+j\omega_2 t_2} dt_1 dt_2. \tag{1.26}$$

The functions $R_{zz}(t_1, t_2)$ and $S_{zz}(\omega_1, \omega_2)$ exhibit Hermitian symmetry, i.e.,

$$R_{zz}(t_1, t_2) = R^*_{zz}(t_2, t_1) \tag{1.27}$$

and

$$S_{zz}(\omega_1, \omega_2) = S^*_{zz}(\omega_2, \omega_1). \tag{1.28}$$

However, there is no requirement that $S_{zz}(\omega_1, \omega_2)$ exhibit the conjugate symmetry for positive and negative frequencies, given in Equation 1.6, as is the case for real stochastic processes.

Other properties of real second-order stochastic processes given above carry over to complex processes. Namely, if $\mathbf{H}$ is a LTI system with arbitrary complex impulse response $h(t)$, frequency response $H(\omega)$, and complex input $\mathbf{z}(t)$, then the complex output $\mathbf{w}(t)$ satisfies

$$S_{ww}(\omega_1, \omega_2) = H(\omega_1) H^*(\omega_2) S_{zz}(\omega_1, \omega_2). \tag{1.29}$$

A bandpass circular complex stochastic process is one with finite spectral support in some arbitrary frequency band B.

Complex stochastic processes undergo a frequency translation when multiplied by a deterministic complex exponential. If $\mathbf{z}(t)$ is circular, then

$$\mathbf{w}(t) = e^{j\omega_c t} \mathbf{z}(t) \tag{1.30}$$

is also circular, and has bifrequency energy spectral density function

$$S_{ww}(\omega_1, \omega_2) = S_{zz}(\omega_1 - \omega_c, \omega_2 - \omega_c). \tag{1.31}$$

1.2.4 Complex Representations of Finite-Energy Second-Order Stochastic Processes

Let $\mathbf{s}(t)$ be a bandpass finite-energy second-order stochastic process, as defined in Section 1.2.2. The complex representation of $\mathbf{s}(t)$ is found by the same down-conversion and filtering operation described for deterministic signals:

$$\mathbf{g}(t) = \mathrm{LPF}\{2\mathbf{s}(t) e^{-j\omega_c t}\}. \tag{1.32}$$

The lowpass filter (LPF) in Equation 1.32 is an ideal filter that passes the baseband components of the frequency-shifted signal, and attenuates the components centered at frequency $-2\omega_c$.

The inverse operation for Equation 1.32 is given by

$$\hat{\mathbf{s}}(t) = \mathrm{Re}\{\mathbf{g}(t) e^{j\omega_c t}\}. \tag{1.33}$$

Because the operation in Equation 1.32 involves the integral of a stochastic process, which we define using mean-square stochastic convergence, we cannot say that $\mathbf{s}(t)$ is identically equal to $\hat{s}(t)$ in the manner that we do for deterministic signals. However, it can be shown that $\mathbf{s}(t)$ and $\hat{s}(t)$ are equivalent in the mean-square sense, that is,

$$E\{(\mathbf{s}(t) - \hat{\mathbf{s}}(t))^2\} = 0, \quad \text{all } t. \tag{1.34}$$

With this interpretation, we say that $\mathbf{g}(t)$ is the unique complex envelope representation for $\mathbf{s}(t)$.

The assumption of circularity of the complex representation is widespread in many signal processing applications. There is an equivalent condition which can be placed on the real bandpass signal that guarantees its complex representation has this circularity property. This condition can be found indirectly by starting with a circular $\mathbf{g}(t)$ and looking at the $\mathbf{s}(t)$ which results.

Let $\mathbf{g}(t)$ be an arbitrary lowpass circular complex finite-energy second-order stochastic process. The frequency-shifted version of this process is

$$\mathbf{p}(t) = \mathbf{g}(t)\mathrm{e}^{+j\omega_c t} \tag{1.35}$$

and the real part of this is

$$\mathbf{s}(t) = \frac{1}{2}(\mathbf{p}(t) + \mathbf{p}^\star(t)). \tag{1.36}$$

By the definition of circularity, $\mathbf{p}(t)$ and $\mathbf{p}^\star(t)$ are orthogonal processes ($E\{\mathbf{p}(t_1)(\mathbf{p}^\star(t_2))^\star = 0\}$) and from this we have

$$\begin{aligned} S_{ss}(\omega_1, \omega_2) &= \frac{1}{4}(S_{pp}(\omega_1, \omega_2) + S_{p^\star p^\star}(\omega_1, \omega_2)) \\ &= \frac{1}{4}(S_{gg}(\omega_1 - \omega_c, \omega_2 - \omega_c) + S^\star_{gg}(-\omega_1 - \omega_c, -\omega_2 - \omega_c)). \end{aligned} \tag{1.37}$$

Since $\mathbf{g}(t)$ is a baseband signal, the first term in Equation 1.37 has spectral support in the first quadrant in the (ω_1, ω_2) plane, where both ω_1 and ω_2 are positive, and the second term has spectral support only for both frequencies negative. This situation is illustrated in Figure 1.6.

It has been shown that a necessary condition for $\mathbf{s}(t)$ to have a circular complex envelope representation is that it have spectral support only in the first and third quadrants of the (ω_1, ω_2) plane. This condition is also sufficient: if $\mathbf{g}(t)$ is not circular, then the $\mathbf{s}(t)$ which results from the operation in Equation 1.33 will have nonzero spectral components in the second and fourth quadrants of the (ω_1, ω_2) plane, and this contradicts the mean-square equivalence of $\mathbf{s}(t)$ and $\hat{\mathbf{s}}(t)$.

An interesting class of processes with spectral support only in the first and third quadrants is the class of processes whose autocorrelation function is separable in the following way:

$$R_{ss}(t_1, t_2) = R_1(t_1 - t_2)R_2\left(\frac{t_1 + t_2}{2}\right). \tag{1.38}$$

For these processes, the bifrequency energy spectral density separates in a like manner:

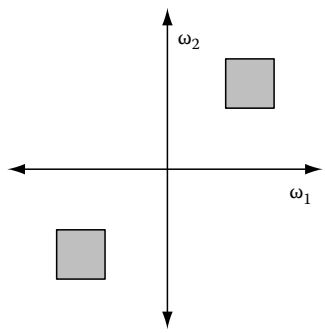

FIGURE 1.6 Spectral support for bandpass process with circular complex representation.

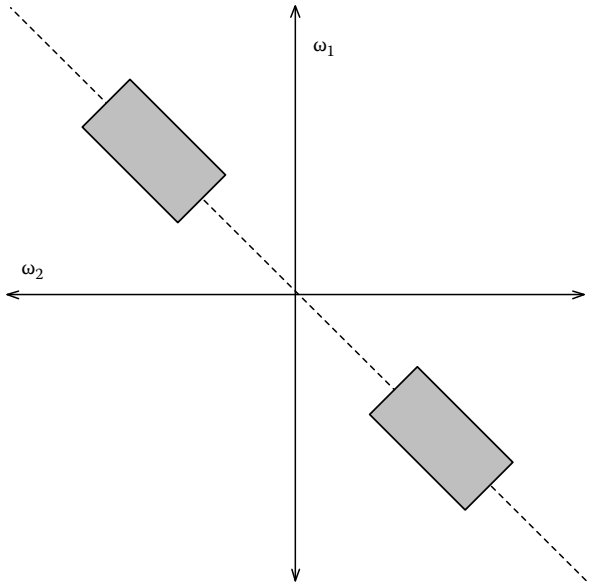

FIGURE 1.7 Spectral support for bandpass process with separable autocorrelation.

$$S_{ss}(\omega_1, \omega_2) = S_1(\omega_1 - \omega_2) S_2\left(\frac{\omega_1 + \omega_2}{2}\right). \tag{1.39}$$

In fact, S_1 is the Fourier transform of R_2 and vice versa. If S_1 is a lowpass function, and S_2 is a bandpass function, then the resulting product has spectral support illustrated in Figure 1.7.

The assumption of circularity in the complex representation can often be physically motivated. For example, in a radar system, if the reflected electromagnetic wave undergoes a phase shift, or if the reflector position cannot be resolved to less than a wavelength, or if the reflection is due to a sum of reflections at slightly different path lengths, then the absolute phase of the return signal is considered random and uniformly distributed. Usually it is not the absolute phase of the received signal which is of interest; rather, it is the "relative phase" of the signal value at two different points in time, or of two different signals at the same instance in time. In many radar systems, particularly those used for direction-of-arrival estimation or delay-Doppler imaging, this relative phase is central to the signal processing objective.

1.2.5 Finite-Power Stochastic Processes

The second major class of second-order processes we wish to consider is the class of finite-power signals. A finite-power signal $\mathbf{x}(t)$ as one whose mean-square value exists, as in Equation 1.4, but whose total energy, as defined in Equation 1.12, is infinite. Furthermore, we require that the time-averaged mean-square value, given by

$$P_x = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} R_{xx}(t, t) \mathrm{d}t, \tag{1.40}$$

exist and be finite. P_x is called the "power" of the process $\mathbf{x}(t)$.

The most commonly invoked stochastic process of this type in communications and signal processing is the "wide-sense-stationary" (w.s.s.) process, one whose autocorrelation function $R_{xx}(t_1, t_2)$

is a function of the time difference $t_1 - t_2$ only. In this case, the mean-square value is constant and is equal to the average power. Such a process is used to model a communication signal that transmits for a long period of time, and for which the beginning and end of transmission are considered unimportant.

A w.s.s. process may be considered to be the limiting case of a particular type of finite-energy process, namely a process with separable autocorrelation as described by Equations 1.38 and 1.39. If in Equation 1.38 the function $R_2\left(\frac{t_1+t_2}{2}\right)$ is equal to a constant, then the process is w.s.s. with second-order properties determined by the function $R_1(t_1 - t_2)$. The bifrequency energy spectral density function is

$$S_{xx}(\omega_1, \omega_2) = 2\pi\delta(\omega_1 - \omega_2)S_2\left(\frac{\omega_1 + \omega_2}{2}\right) \tag{1.41}$$

where

$$S_2(\omega) = \int_{-\infty}^{\infty} R_1(\tau)e^{-j\omega\tau}d\tau. \tag{1.42}$$

This last pair of equations motivates us to describe the second-order properties of $\mathbf{x}(t)$ with functions of one argument instead of two, namely the autocorrelation function $R_{xx}(\tau)$ and its Fourier transform $S_{xx}(\omega)$, known as the power spectral density. From basic Fourier transform properties we have

$$P_x = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{xx}(\omega)d\omega. \tag{1.43}$$

If w.s.s. $\mathbf{x}(t)$ is the input to a LTI system with frequency response $H(\omega)$ and output $\mathbf{y}(t)$, then it is not difficult to show that

1. $\mathbf{y}(t)$ is w.s.s.
2. $S_{yy}(\omega) = |H(\omega)|^2 S_{xx}(\omega)$.

These last results, combined with Equation 1.43, lead to a natural interpretation of the power spectral density function. If $\mathbf{x}(t)$ is the input to an ideal bandpass filter with passband B, then the total power of the filter output is

$$P_y = \frac{1}{2\pi}\int_{B} S_x(\omega)d\omega. \tag{1.44}$$

This shows how the total power in the process $\mathbf{x}(t)$ can be attributed to components in different spectral bands.

1.2.6 Complex Wide-Sense-Stationary Processes

Two real stochastic processes $\mathbf{x}(t)$ and $\mathbf{y}(t)$, defined on a common probability space, are said to be jointly w.s.s. if:

1. Both $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are w.s.s.
2. The cross-correlation $R_{xy}(t_1, t_2) = E\{\mathbf{x}(t_1)\mathbf{y}(t_2)\}$ is a function of $t_1 - t_2$ only.

For jointly w.s.s. processes, the cross-correlation function is normally written with a single argument, e.g., $R_{xy}(\tau)$, with $\tau = t_1 - t_2$. From the definition we see that

$$R_{xy}(\tau) = R_{yx}(-\tau). \tag{1.45}$$

A complex w.s.s. stochastic process $\mathbf{z}(t)$ is one that can be written

$$\mathbf{z}(t) = \mathbf{x}(t) + j\mathbf{y}(t) \tag{1.46}$$

where $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are jointly w.s.s. A "circular" complex w.s.s. process is one in which

$$R_{xx}(\tau) = R_{yy}(\tau) \tag{1.47}$$

and

$$R_{xy}(\tau) = -R_{yx}(\tau), \quad \text{all } \tau. \tag{1.48}$$

The reader is cautioned not to confuse the meanings of Equations 1.45 and 1.48.

For circular complex w.s.s. processes, it is easy to show that

$$E\{\mathbf{z}(t_1)\mathbf{z}(t_2)\} = 0, \quad \text{all } t_1, t_2, \tag{1.49}$$

and therefore the function

$$\begin{aligned} R_{zz}(t_1, t_2) &= E\{\mathbf{z}(t_1)\mathbf{z}^{\star}(t_2)\} \\ &= 2R_{xx}(t_1, t_2) + 2jR_{yx}(t_1, t_2) \end{aligned} \tag{1.50}$$

defines all the second-order properties of $\mathbf{z}(t)$. All the quantities involved in Equation 1.50 are functions of $\tau = t_1 - t_2$ only, and thus the single-argument function $R_{zz}(\tau)$ is defined as the autocorrelation function for $\mathbf{z}(t)$.

The power spectral density for $\mathbf{z}(t)$ is

$$S_{zz}(\omega) = \int_{-\infty}^{\infty} R_{zz}(\tau)\mathrm{e}^{-j\omega\tau}\mathrm{d}\tau. \tag{1.51}$$

$R_{zz}(\tau)$ exhibits conjugate symmetry ($R_{zz}(\tau) = R_{zz}^{*}(-\tau)$); $S_{zz}(\omega)$ is nonnegative but otherwise has no symmetry constraints.

If $\mathbf{z}(t)$ is the input to a complex LTI system with frequency response $H(\omega)$, then the output process $\mathbf{w}(t)$ is wide-sense-stationarity with power spectral density

$$S_{ww}(\omega) = |H(\omega)|^2 S_{zz}(\omega). \tag{1.52}$$

A bandpass w.s.s. process is one with finite (possible asymmetric) support in frequency.

If $\mathbf{z}(t)$ is a circular w.s.s. process, then

$$\mathbf{w}(t) = \mathrm{e}^{j\omega_c t}\mathbf{z}(t) \tag{1.53}$$

is also circular, and has power spectral density

$$S_{ww}(\omega) = S_{zz}(\omega - \omega_c). \tag{1.54}$$

1.2.7 Complex Representations of Real Wide-Sense-Stationary Signals

Let $\mathbf{s}(t)$ be a real bandpass w.s.s. stochastic process. The complex representation for $\mathbf{s}(t)$ is given by the now-familiar expression

$$\mathbf{g}(t) = \text{LPF}\{2\mathbf{s}(t)e^{-j\omega_c t}\} \tag{1.55}$$

with inverse relationship

$$\hat{\mathbf{s}}(t) = \text{Re}\{\mathbf{g}(t)e^{j\omega_c t}\}. \tag{1.56}$$

In Equations 1.55 and 1.56, ω_c is the center frequency for the passband of $\mathbf{s}(t)$, and the LPF has bandwidth greater than that of $\mathbf{s}(t)$ but much less than $2\omega_c$. $\mathbf{s}(t)$ and $\hat{\mathbf{s}}(t)$ are equivalent in the mean-square sense, implying that $\mathbf{g}(t)$ is the unique complex envelope representation for $\mathbf{s}(t)$.

For arbitrary real w.s.s. $\mathbf{s}(t)$, the circularity of the complex representation comes without any additional conditions like the ones imposed for finite-energy signals. If w.s.s. $\mathbf{s}(t)$ is the input to a quadrature demodulator, then the output signals $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are jointly w.s.s., and the complex process

$$\mathbf{g}(t) = \mathbf{x}(t) + j\mathbf{y}(t) \tag{1.57}$$

is circular. There are various ways of showing this, with the simplest probably being a proof by contradiction. If $\mathbf{g}(t)$ is a complex process that is not circular, then the process $\text{Re}\{\mathbf{g}(t)e^{j\omega_c t}\}$ can be shown to have an autocorrelation function with nonzero terms which are a function of $t_1 + t_2$, and thus it cannot be w.s.s.

Communication signals are often modeled as w.s.s. stochastic processes. The stationarity results from the fact that the carrier phase, as seen at the receiver, is unknown and considered random, due to lack of knowledge about the transmitter and path length. This in turn leads to a circularity assumption on the complex modulation.

In many communication and surveillance systems, the quadrature demodulator is an actual electronic subsystem which generates a pair of signals interpreted directly as a complex representation of a bandpass signal. Often these signals are sampled, providing complex digital data for further digital signal processing. In array signal processing, there are multiple such receivers, one behind each sensor or antenna in a multisensor system. Data from an array of receivers is then modeled as a "vector" of complex random variables. In the next section, we consider multivariate distributions for such complex data.

1.3 The Multivariate Complex Gaussian Density Function

The discussions of Section 1.2 centered on the second-order (correlation) properties of real and complex stochastic processes, but to this point nothing has been said about joint probability distributions for these processes. In this section, we consider the distribution of samples from a complex process in which the real and imaginary parts are Gaussian distributed. The key concept of this section is that the assumption of circularity on a complex stochastic process (or any collection of complex random variables) leads to a compact form of the density function which can be written directly as a function of a complex argument z rather than its real and imaginary parts.

From a data processing point-of-view, a collection of N complex numbers is simply a collection of $2N$ real numbers, with a certain mathematical significance attached to the N numbers we call the "real parts"

and the other N numbers we call the "imaginary parts." Likewise, a collection of N complex random variables is really just a collection of $2N$ real random variables with some joint distribution in R^{2N}. Because these random variables have an interpretation as real and imaginary parts of some complex numbers, and because the $2N$-dimensional distribution may have certain symmetries such as those resulting from circularity, it is often natural and intuitive to express joint densities and distributions using a notation which makes explicit the complex nature of the quantities involved. In this section we develop such a density for the case where the random variables have a Gaussian distribution and are samples of a circular complex stochastic process.

Let $\mathbf{z}_i, i = 1, \ldots, N$ be a collection of complex numbers that we wish to model probabilistically. Write

$$\mathbf{z}_i = \mathbf{x}_i + j\mathbf{y}_i \tag{1.58}$$

and consider the vector of numbers $[\mathbf{x}_1, \mathbf{y}_1, \ldots, \mathbf{x}_N, \mathbf{y}_N]^{\mathrm{T}}$ as a set of $2N$ random variables with a distribution over $\mathbb{R}^{2N}$. Suppose further that the vector $[\mathbf{x}_1, \mathbf{y}_1, \ldots, \mathbf{x}_N, \mathbf{y}_N]^{\mathrm{T}}$ is subject to the usual multivariate Gaussian distribution with $2N \times 1$ mean vector μ and $2N \times 2N$ covariance matrix $\mathbf{R}$. For compactness, denote the entire random vector with the symbol $\mathbf{x}$. The density function is

$$f_{\mathrm{x}}(x) = (2\pi)^{\frac{-2N}{2}} (\det \mathbf{R})^{\frac{-1}{2}} \mathrm{e}^{-\frac{x^{\mathrm{T}}\mathbf{R}^{-1}x}{2}}. \tag{1.59}$$

We seek a way of expressing the density function of Equation 1.59 directly in terms of the complex variable z, i.e., a density of the form $f_z(z)$. In so doing it is important to keep in mind what such a density represents. $f_{\mathrm{z}}(z)$ will be a nonnegative real-valued function $f : \mathbb{C}^N \to \mathbb{R}^+$, with the property that

$$\int_{\mathbb{C}^N} f_{\mathrm{z}}(z)\mathrm{d}z = 1. \tag{1.60}$$

The probability that $\mathbf{z} \in A$, where A is some subset of $\mathbb{C}^N$, is given by

$$P(A) = \int_A f_{\mathrm{z}}(z)\mathrm{d}z. \tag{1.61}$$

The differential element $\mathrm{d}z$ is understood to be

$$\mathrm{d}z = \mathrm{d}x_1 \mathrm{d}y_1 \mathrm{d}x_2 \mathrm{d}y_2 \ldots \mathrm{d}x_N \mathrm{d}y_N. \tag{1.62}$$

The most general form of the complex multivariate Gaussian density is in fact given by Equation 1.59, and further simplification requires further assumptions. Circularity of the underlying complex process is one such key assumption, and it is now imposed. To keep the following development simple, it is assumed that the mean vector μ is 0. The results for nonzero μ are not difficult to obtain by extension.

Consider the four real random variables $\mathbf{x}_i$, $\mathbf{y}_i$, $\mathbf{x}_k$, $\mathbf{y}_k$. If these numbers represent the samples of a circular complex stochastic process, then we can express the 4×4 covariance as

$$E\left\{ \begin{bmatrix} \mathbf{x}_i \\ \mathbf{y}_i \\ \mathbf{x}_k \\ \mathbf{y}_k \end{bmatrix} [\mathbf{x}_i \mathbf{y}_i \mathbf{x}_k \mathbf{y}_k] \right\} = \frac{1}{2} \left[\begin{array}{cc|cc} \alpha_{ii} & 0 & \alpha_{ik} & -\beta_{ik} \\ 0 & \alpha_{ii} & \beta_{ik} & \alpha_{ik} \\ \hline \alpha_{ki} & -\beta_{ki} & \alpha_{kk} & 0 \\ \beta_{ki} & \alpha_{ki} & 0 & \alpha_{kk} \end{array}\right], \tag{1.63}$$

where

$$\alpha_{ik} = 2E\{\mathbf{x}_i\mathbf{x}_k\} = 2E\{\mathbf{y}_i\mathbf{y}_k\} \tag{1.64}$$

and

$$\beta_{ik} = -2E\{\mathbf{x}_i\mathbf{y}_k\} = +2E\{\mathbf{x}_k\mathbf{y}_i\}. \tag{1.65}$$

Extending this to the full $2N \times 2N$ covariance matrix $\mathbf{R}$, we have

$$\mathbf{R} = \frac{1}{2}\left[\begin{array}{cc|cc|ccccc|cc}
\alpha_{11} & 0 & \alpha_{12} & -\beta_{12} & \cdot & \cdot & \cdot & \cdot & \cdot & \alpha_{1N} & -\beta_{1N} \\
0 & \alpha_{11} & \beta_{12} & \alpha_{12} & \cdot & \cdot & \cdot & \cdot & \cdot & \beta_{1N} & \alpha_{1N} \\
- & - & - & - & - & - & - & - & - & - & - \\
\alpha_{21} & -\beta_{21} & \alpha_{22} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \alpha_{2N} & -\beta_{2N} \\
\beta_{21} & \alpha_{21} & 0 & \alpha_{22} & \cdot & \cdot & \cdot & \cdot & \cdot & \beta_{2N} & \alpha_{2N} \\
- & - & - & - & - & - & - & - & - & - & - \\
\cdot & \cdot & \cdot & \cdot & \cdot & & & & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & & \cdot & & & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & & & \cdot & & \cdot & \cdot & \cdot \\
- & - & - & - & - & - & - & - & - & - & - \\
\alpha_{N1} & -\beta_{N1} & \alpha_{N2} & -\beta_{N2} & \cdot & \cdot & \cdot & \cdot & \cdot & \alpha_{NN} & 0 \\
\beta_{N1} & \alpha_{N1} & \beta_{N2} & \alpha_{N2} & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & \alpha_{NN}
\end{array}\right]. \tag{1.66}$$

The key thing to notice about the matrix in Equation 1.66 is that, because of its special structure, it is completely specified by N^2 real quantities: one for each of the 2×2 diagonal blocks, and two for each of the 2×2 upper off-diagonal blocks. This is in contrast to the $N(2N+1)$ free parameters one finds in an unconstrained $2N \times 2N$ real Hermitian matrix.

Consider now the complex random variables $\mathbf{z}_i$ and $\mathbf{z}_k$. We have that

$$\begin{aligned} E\{\mathbf{z}_i\mathbf{z}_i^*\} &= E\{(\mathbf{x}_i + j\mathbf{y}_i)(\mathbf{x}_i - j\mathbf{y}_i)\} \\ &= E\{\mathbf{x}_i^2 + \mathbf{y}_i^2\} = \alpha_{ii} \end{aligned} \tag{1.67}$$

and

$$\begin{aligned} E\{\mathbf{z}_i\mathbf{z}_k^*\} &= E\{(\mathbf{x}_i + j\mathbf{y}_i)(\mathbf{x}_k - j\mathbf{y}_k)\} \\ &= E\{\mathbf{x}_i\mathbf{x}_k + \mathbf{y}_i\mathbf{y}_k - j\mathbf{x}_k\mathbf{y}_i + j\mathbf{x}_i\mathbf{y}_k\} \\ &= \alpha_{ik} + j\beta_{ik}. \end{aligned} \tag{1.68}$$

Similarly

$$E\{\mathbf{z}_k\mathbf{z}_i^*\} = \alpha_{ik} - j\beta_{ik} \tag{1.69}$$

and

$$E\{\mathbf{z}_k\mathbf{z}_k^*\} = \alpha_{kk}. \tag{1.70}$$

Using Equations 1.66 through 1.70, it is possible to write the following $N \times N$ complex Hermitian matrix:

$$E\{\mathbf{z}\mathbf{z}^{\mathrm{H}}\} = \begin{bmatrix} \alpha_{11} & | & \alpha_{12} + j\beta_{12} & | & \cdots & | & \alpha_{1N} + j\beta_{1N} \\ --- & - & --- & - & --- & - & --- \\ \alpha_{21} + j\beta_{21} & | & \alpha_{22} & | & \cdots & | & \alpha_{2N} + j\beta_{2N} \\ --- & - & --- & - & --- & - & --- \\ \cdot & | & \cdot & | & \cdot & | & \cdot \\ \cdot & | & \cdot & | & \cdot & | & \cdot \\ \cdot & | & \cdot & | & \cdot & | & \cdot \\ --- & - & --- & - & --- & - & --- \\ \alpha_{N1} + j\beta_{N1} & | & \alpha_{N2} + j\beta_{N2} & | & \cdots & | & \alpha_{NN} \end{bmatrix}. \tag{1.71}$$

Note that this complex matrix has exactly the same N^2 free parameters as did the $2N \times 2N$ real matrix $\mathbf{R}$ in Equation 1.66, and thus it tells us everything there is to know about the joint distribution of the real and imaginary components of $\mathbf{z}$. Under the symmetry constraints imposed on $\mathbf{R}$, we can define

$$\mathbf{C} = E\{\mathbf{z}\mathbf{z}^{\mathrm{H}}\} \tag{1.72}$$

and call this matrix the covariance matrix for $\mathbf{z}$. In the 0-mean Gaussian case, this matrix parameter uniquely identifies the multivariate distribution for $\mathbf{z}$.

The derivation of the density function $f_z(z)$ rests on a set of relationships between the $2N \times 1$ real vector $\mathbf{x}$, and its $N \times 1$ complex counterpart $\mathbf{z}$. We say that $\mathbf{x}$ and $\mathbf{z}$ are "isomorphic" to one another, and denote this with the symbol

$$\mathbf{z} \approx \mathbf{x}. \tag{1.73}$$

Likewise we say that the $2N \times 2N$ real matrix $\mathbf{R}$, given in Equation 1.66, and the $N \times N$ complex matrix $\mathbf{C}$, given in Equation 1.71 are isomorphic to one another, or

$$\mathbf{C} \approx \mathbf{R}. \tag{1.74}$$

The development of the complex Gaussian density function $f_z(z)$ is based on three claims based on these isomorphisms.

Proposition 1.1. If $\mathbf{z} \approx \mathbf{x}$, and $\mathbf{R} \approx \mathbf{C}$, then

$$\mathbf{x}^{\mathrm{T}}(2\mathbf{R})\mathbf{x} = \mathbf{z}^{\mathrm{H}}\mathbf{C}\mathbf{z}. \tag{1.75}$$

Proposition 1.2. If $\mathbf{R} \approx \mathbf{C}$, then

$$\frac{1}{4}\mathbf{R}^{-1} \approx \mathbf{C}^{-1}. \tag{1.76}$$

Proposition 1.3. If $\mathbf{R} \approx \mathbf{C}$, then

$$\det \mathbf{R} = |\det \mathbf{C}|^2 \left(\frac{1}{2}\right)^{2N}. \tag{1.77}$$

The density function $f_z(z)$ is found by substituting the results from Propositions 1.1 through 1.3 directly into the density function $f_x(x)$. This is possible because the mapping from $\mathbf{z}$ to $\mathbf{x}$ is one-to-one and onto, and the Jacobian is 1 [see Equation 1.62]. We have

$$\begin{aligned} f_z(z) &= (2\pi)^{\frac{-2N}{2}} (\det \mathbf{R})^{\frac{-1}{2}} e^{-\frac{x^T \mathbf{R}^{-1} x}{2}} \\ &= \left(\frac{1}{2}\right)^{-N} (2\pi)^{-N} (\det \mathbf{C})^{-1} e^{-z^H \mathbf{C}^{-1} z} \qquad (1.78) \\ &= \pi^{-N} (\det \mathbf{C})^{-1} e^{-z^H \mathbf{C}^{-1} z}. \qquad (1.79) \end{aligned}$$

At this point it is straightforward to introduce a nonzero mean μ, which is the complex vector isomorphic to the mean of the real random vector $\mathbf{x}$. The resulting density is

$$f_z(z) = \pi^{-N} (\det \mathbf{C})^{-1} e^{-(z-\mu)^H \mathbf{C}^{-1} (z-\mu)}. \qquad (1.80)$$

The density function in Equation 1.80 is commonly referred to as the "complex Gaussian density function," although in truth one could be more general and have an arbitrary $2N$-dimension Gaussian distribution on the real and imaginary components of $\mathbf{z}$. It is important to recognize that the use of Equation 1.80 implies those symmetries in the real covariance of $\mathbf{x}$ implied by circularity of the underlying complex process. This symmetry is expressed by some authors in the equation

$$E\{\mathbf{z}\mathbf{z}^T\} = 0 \qquad (1.81)$$

where the superscript "T" indicates transposition without complex conjugation. This comes directly from Equations 1.24 and 1.49.

For many, the functional form of the complex Gaussian density in Equation 1.80 is actually simpler and cleaner than its N-dimensional real counterpart, due to elimination of the various factors of 2 which complicate it. This density is the starting point for virtually all of the multivariate analysis of complex data seen in the current signal and array processing literature.

1.4 Related Distributions

In many problems of interest in statistical signal processing, the raw data may be complex and subject to a complex Gaussian distribution described in the density function in Equation 1.80. The processing may take the form of the computation of a test statistic for use in a hypothesis test. The density functions for these test statistics are then used to determine probabilities of false alarm and/or detection. Thus, it is worthwhile to study certain distributions that are closely related to the complex Gaussian in this way.

In this section we will describe and give the functional form for four densities related to the complex Gaussian: the complex χ^2, the complex F, the complex β, and the complex t. Only the "central" versions of these distributions will be given, i.e., those based on 0-mean Gaussian data. The central distributions are usually associated with the null hypothesis in a detection problem and are used to compute probabilities of false alarm. The noncentral densities, used in computing probabilities of detection, do not exist in closed form but can be easily tabulated.

1.4.1 Complex Chi-Squared Distribution

One very common type of detection problem in radar problems is the "signal present" vs. "signal absent" decision problem. Often under the "signal absent" hypothesis, the data is zero-mean complex Gaussian, with known covariance, whereas under the "signal present" hypothesis the mean is nonzero, but perhaps

unknown or subject to some uncertainty. A common test under these circumstances is to compute the sum of squared magnitudes of the data points (after prewhitening, if appropriate) and compare this to a threshold. The resulting test statistic has a χ^2-squared distribution.

Let $\mathbf{z}_1 \ldots \mathbf{z}_N$ be N complex Gaussian random variables, independent and identically distributed with mean 0 and variance 1 (meaning that the covariance matrix for the $\mathbf{z}$ vector is $\mathbf{I}$). Define the real nonnegative random variable $\mathbf{q}$ according to

$$\mathbf{q} = \sum_{i}^{N} |\mathbf{z}_i|^2. \tag{1.82}$$

Then the density function for $\mathbf{q}$ is given by

$$f_{\mathrm{q}}(q) = \frac{1}{(N-1)!} q^{N-1} \mathrm{e}^{-q} U(q). \tag{1.83}$$

To establish this result, show that the density function for $|\mathbf{z}_i|^2$ is a simple exponential. Equation 1.83 is the N-fold convolution of this exponential density function with itself.

We often say that $\mathbf{q}$ is χ^2 with N complex degrees of freedom. A "complex degree of freedom" is like two real degrees of freedom. Note, however, that Equation 1.83 is not the usual χ^2 density function with $2N$ degrees of freedom. Each of the real variables going into the computation of $\mathbf{q}$ has variance $\frac{1}{2}$, not 1. $f_{\mathrm{q}}(q)$ is a gamma density with an integer parameter N, and, like the complex Gaussian density in Equation 1.60, it is cleaner and simpler than its real counterpart.

1.4.2 Complex *F*-Distribution

In some "signal present" vs. "signal absent" problems, the variance or covariance of the noise is not known under the null hypothesis, and must be estimated from some auxiliary data. Then the test statistic becomes the ratio of the sum of square magnitudes of the test data to the sum of square magnitudes of the auxiliary data. The resulting test statistic is subject to a particular form of the F-distribution.

Let $\mathbf{q}_1$ and $\mathbf{q}_2$ be two independent random variables subject to the χ^2 distribution with N and M complex degrees of freedom, respectively. Define the real, nonnegative random variable $\mathbf{f}$ according to

$$\mathbf{f} = \frac{\mathbf{q}_1}{\mathbf{q}_2}. \tag{1.84}$$

The density function for $\mathbf{f}$ is

$$f_{\mathrm{f}}(f) = \frac{(N+M-1)!}{(N-1)!(M-1)!} \frac{f^{N-1}}{(1+f)^{N+M}} U(f). \tag{1.85}$$

We say that $\mathbf{f}$ is subject to an F-distribution with N and M complex degrees of freedom.

1.4.3 Complex Beta Distribution

An F-distributed random variable can be transformed in such a way that the resulting density has finite support. The random variable $\mathbf{b}$, defined by

$$\mathbf{b} = \frac{1}{(1+\mathbf{f})}, \tag{1.86}$$

where **f** is an *F*-distributed random variable, has this property. The density function is given by

$$f_{\mathrm{b}}(b) = \frac{(N+M-1)!}{(N-1)!(M-1)!} b^{M-1}(1-b)^{N-1} \tag{1.87}$$

on the interval $0 \leq b \leq 1$, and is 0 elsewhere.

The random variable **b** is said to be beta-distributed, with *N* and *M* complex degrees of freedom.

1.4.4 Complex Student-*t* Distribution

In the "signal present" vs. "signal absent" problem, if the signal is known exactly (including phase) then the optimal detector is a prewhitener followed by a matched filter. The resulting test statistic is complex Gaussian, and the detector partitions the complex plane into two half-planes which become the decision regions for the two hypotheses. Now it may be that the signal is known, but the variance of the noise is not. In this case, the Gaussian test statistic must be scaled by an estimate of the standard deviation, obtained as before from zero-mean auxiliary data. In this case the test statistic is said to have a complex *t* (or Student-*t*) distribution. Of the four distributions discussed in this section, this is the only one in which the random variables themselves are complex: the χ^2, *F*, and β distributions all describe real random variables functionally dependent on complex Gaussians.

Let **z** and **q** be independent scalar random variables. **z** is complex Gaussian with mean 0 and variance 1, and **q** is χ^2 with *N* complex degrees of freedom. Define the random variable **t** according to

$$\mathbf{t} = \frac{\mathbf{z}}{\sqrt{\mathbf{q}/N}}. \tag{1.88}$$

The density of **t** is then given by

$$f_{\mathrm{t}}(t) = \frac{1}{\pi\left(1+\frac{|t|^2}{N}\right)^{N+1}}. \tag{1.89}$$

This density is said to be "heavy-tailed" relative to the Gaussian, and this is a result in the uncertainty in the estimate of the standard deviation. Note that as $N \to \infty$, the denominator Equation 1.88 approaches 1 (i.e., the estimate of the standard deviation approaches truth) and thus $f_{\mathrm{t}}(t)$ approaches the Gaussian density $\pi^{-1}\mathrm{e}^{-|t|^2}$ as expected.

1.5 Conclusion

In this chapter, we have outlined a basic theory of complex random variables and stochastic processes as they most often appear in statistical signal and array processing problems. The properties of complex representations for real bandpass signals were emphasized, since this is the most common application in electrical engineering where complex data appear. Models for both finite-energy signals, such as radar pulses, and finite-power signals, such as communication signals, were developed. The key notion of circularity of complex stochastic processes was explored, along with the conditions that a real stochastic process must satisfy in order for it to have a circular complex representation. The complex multivariate Gaussian distribution was developed, again building on the circularity of the underlying complex stochastic process. Finally, related distributions which often appear in statistical inference problems with complex Gaussian data were introduced.

The general topic of random variables and stochastic processes is fundamental to modern signal processing, and many good textbooks are available. Those by Papoulis [2], Leon-Garcia [3], and Melsa

and Sage [4] are recommended. The original short paper deriving the complex multivariate Gaussian density function is by Wooding [5]; another derivation and related statistical analysis is given in Goodman [6], whose name is more often cited in connection with complex random variables. The monograph by Miller [7] has a mathematical flavor, and covers complex stochastic processes, stochastic differential equations, parameter estimation, and least-squares problems. The paper by Neeser and Massey [8] treats circular (which they call "proper") complex stochastic processes and their application in information theory. There is a good discussion of complex random variables in Kay [9], which includes Cramer–Rao lower bounds and optimization of functions of complex variables. Kelly and Forsythe [10] is an advanced treatment of inference problems for complex multivariate data, and contains a number of appendices with valuable background information, including one on distributions related to the complex Gaussian.

References

1. Loeve, M., *Probability Theory*, D. Van Nostrand Company, New York, 1963.
2. Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 3rd edn., McGraw-Hill, New York, 1991.
3. Leon-Garcia, A., *Probability and Random Processes for Electrical Engineering*, 2nd edn., Addison-Wesley, Reading, MA, 1994.
4. Melsa, J. and Sage, A., *An Introduction to Probability and Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
5. Wooding, R., The multivariate distribution of complex normal variables, *Biometrika*, 43, 212–215, 1956.
6. Goodman, N., Statistical analysis based on a certain multivariate complex Gaussian distribution, *Ann. Math. Stat*., 34, 152–177, 1963.
7. Miller, K., *Complex Stochastic Processes*, Addison-Wesley, Reading, MA, 1974.
8. Neeser, F. and Massey, J., Proper complex random processes with applications to information theory, *IEEE Trans. Inform. Theory*, 39(4), 1293–1302, July 1993.
9. Kay, S., *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
10. Kelly, E. and Forsythe, K., Adaptive detection and parameter estimation for multidimensional signal models, MIT Lincoln Laboratory Technical Report 848, April 1989.

2

Beamforming Techniques for Spatial Filtering

Barry Van Veen
University of Wisconsin

Kevin M. Buckley
Villanova University

2.1 Introduction

Systems designed to receive spatially propagating signals often encounter the presence of interference signals. If the desired signal and interferers occupy the same temporal frequency band, then temporal filtering cannot be used to separate signal from interference. However, desired and interfering signals often originate from different spatial locations. This spatial separation can be exploited to separate signal from interference using a spatial filter at the receiver.

A beamformer is a processor used in conjunction with an array of sensors to provide a versatile form of spatial filtering. The term "beamforming" derives from the fact that early spatial filters were designed to form pencil beams (see polar plot in Figure 2.5c) in order to receive a signal radiating from a specific location and attenuate signals from other locations. "Forming beams" seems to indicate radiation of energy; however, beamforming is applicable to either radiation or reception of energy. In this section we discuss the formation of beams for reception, providing an overview of beamforming from a signal processing perspective. Data-independent, statistically optimum, adaptive, and partially adaptive beamforming are discussed.

Implementing a temporal filter requires processing of data collected over a temporal aperture. Similarly, implementing a spatial filter requires processing of data collected over a spatial aperture. A single sensor

such as an antenna, sonar transducer, or microphone collects impinging energy over a continuous aperture, providing spatial filtering by summing coherently waves that are in phase across the aperture while destructively combining waves that are not. An array of sensors provides a discrete sampling across its aperture. When the spatial sampling is discrete, the processor that performs the spatial filtering is termed a beamformer. Typically a beamformer linearly combines the spatially sampled time series from each sensor to obtain a scalar output time series in the same manner that an FIR filter linearly combines temporally sampled data. Two principal advantages of spatial sampling with an array of sensors are discussed in the following.

Spatial discrimination capability depends on the size of the spatial aperture; as the aperture increases, discrimination improves. The absolute aperture size is not important, rather its size in wavelengths is the critical parameter. A single physical antenna (continuous spatial aperture) capable of providing the requisite discrimination is often practical for high-frequency signals because the wavelength is short. However, when low-frequency signals are of interest, an array of sensors can often synthesize a much larger spatial aperture than that practical with a single physical antenna.

A second very significant advantage of using an array of sensors, relevant at any wavelength, is the spatial filtering versatility offered by discrete sampling. In many application areas, it is necessary to change the spatial filtering function in real time to maintain effective suppression of interfering signals. This change is easily implemented in a discretely sampled system by changing the way in which the beamformer linearly combines the sensor data. Changing the spatial filtering function of a continuous aperture antenna is impractical.

This section begins with the definition of basic terminology, notation, and concepts. Succeeding sections cover data-independent, statistically optimum, adaptive, and partially adaptive beamforming. We then conclude with a summary.

Throughout this section we use methods and techniques from FIR filtering to provide insight into various aspects of spatial filtering with beamformer. However, in some ways beamforming differs significantly from FIR filtering. For example, in beamforming a source of energy has several parameters that can be of interest: range, azimuth and elevation angles, polarization, and temporal frequency content. Different signals are often mutually correlated as a result of multipath propagation. The spatial sampling is often nonuniform and multidimensional. Uncertainty must often be included in characterization of individual sensor response and location, motivating development of robust beamforming techniques. These differences indicate that beamforming represents a more general problem than FIR filtering and, as a result, more general design procedures and processing structures are common.

2.2 Basic Terminology and Concepts

In this section we introduce terminology and concepts employed throughout. We begin by defining the beamforming operation and discussing spatial filtering. Next we introduce second-order statistics of the array data, developing representations for the covariance of the data received at the array and discussing distinctions between narrowband and broadband beamforming. Last, we define various types of beamformers.

2.2.1 Beamforming and Spatial Filtering

Figure 2.1 depicts two beamformers. The first, which samples the propagating wave field in space, is typically used for processing narrowband signals. The output at time k, $y(k)$, is given by a linear combination of the data at the J sensors at time k:

$$y(k) = \sum_{l=1}^{J} w_l^* x_l(k), \tag{2.1}$$

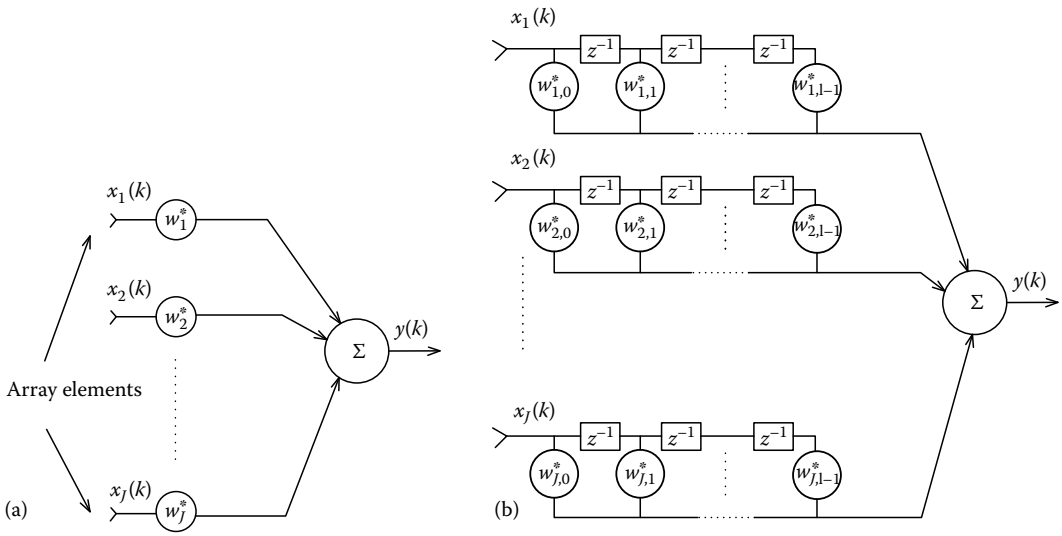

FIGURE 2.1 A beamformer forms a linear combination of the sensor outputs. In (a), sensor outputs are multiplied by complex weights and summed. This beamformer is typically used with narrowband signals. A common broadband beamformer is illustrated in (b).

where $*$ represents complex conjugate. It is conventional to multiply the data by conjugates of the weights to simplify notation. We assume throughout that the data and weights are complex since in many applications a quadrature receiver is used at each sensor to generate in phase and quadrature (I and Q) data. Each sensor is assumed to have any necessary receiver electronics and an A/D converter if beamforming is performed digitally.

The second beamformer in Figure 2.1 samples the propagating wave field in both space and time and is often used when signals of significant frequency extent (broadband) are of interest. The output in this case can be expressed as

$$y(k) = \sum_{l=1}^{J} \sum_{p=0}^{K-1} w_{l,p}^{\star} x_l(k-p), \tag{2.2}$$

where $K-1$ is the number of delays in each of the J sensor channels. If the signal at each sensor is viewed as an input, then a beamformer represents a multi-input single output system.

It is convenient to develop notation that permits us to treat both beamformers in Figure 2.1 simultaneously. Note that Equations 2.1 and 2.2 can be written as

$$y(k) = \mathbf{w}^{\mathrm{H}}\mathbf{x}(k), \tag{2.3}$$

by appropriately defining a weight vector $\mathbf{w}$ and data vector $\mathbf{x}(k)$. We use lower and uppercase boldface to denote vector and matrix quantities, respectively, and let superscript represent Hermitian (complex conjugate) transpose. Vectors are assumed to be column vectors. Assume that $\mathbf{w}$ and $\mathbf{x}(k)$ are N dimensional; this implies that $N=KJ$ when referring to Equation 2.2 and $N=J$ when referring to Equation 2.1. Except for Section 2.5 on adaptive algorithms, we will drop the time index and assume that its presence is understood throughout the remainder of this chapter. Thus, Equation 2.3 is written as $y=\mathbf{w}^{\mathrm{H}}\mathbf{x}$. Many of the techniques described in this section are applicable to continuous time as well as discrete time beamforming.

The frequency response of an FIR filter with tap weights $w_p^\star$, $1 \le p \le J$ and a tap delay of T seconds is given by

$$r(\omega) = \sum_{p=1}^{J} w_p^\star e^{-j\omega T(p-1)}. \tag{2.4}$$

Alternatively

$$r(\omega) = \mathbf{w}^{\mathrm{H}}\mathbf{d}(\omega), \tag{2.5}$$

where

$\mathbf{w}^{\mathrm{H}} = [w_1^\star \; w_2^\star \ldots w_J^\star]$

$\mathbf{d}(\omega) = [1 \; e^{j\omega T} \; e^{j\omega 2T} \cdots e^{j\omega(J-1)T}]^{\mathrm{H}}$

$r(\omega)$ represents the response of the filter* to a complex sinusoid of frequency ω

$\mathbf{d}(\omega)$ is a vector describing the phase of the complex sinusoid at each tap in the FIR filter relative to the tap associated with w_1

Similarly, beamformer response is defined as the amplitude and phase presented to a complex plane wave as a function of location and frequency. Location is, in general, a three-dimensional quantity, but often we are only concerned with one- or two-dimensional direction of arrival (DOA). Throughout the remainder of the section we do not consider range. Figure 2.2 illustrates the manner in which an array of sensors samples a spatially propagating signal. Assume that the signal is a complex plane wave with DOA θ and frequency ω. For convenience let the phase be zero at the first sensor. This implies $x_1(k) = e^{j\omega k}$ and $x_l(k) = e^{j\omega[k - \Delta_l(\theta)]}$, $2 \le l \le J$. $\Delta_l(\theta)$ represents the time delay due to propagation from the first to the lth sensor. Substitution into Equation 2.2 results in the beamformer output

$$y(k) = e^{j\omega k} \sum_{l=1}^{J} \sum_{p=0}^{K-1} w_{l,p}^\star e^{-j\omega[\Delta_l(\theta)+p]} = e^{j\omega k} \; r(\theta\omega), \tag{2.6}$$

where $\Delta_1(\theta) = 0$. $r(\theta, \omega)$ is the beamformer response and can be expressed in vector form as

$$r(\theta, \omega) = \mathbf{w}^{\mathrm{H}}\mathbf{d}(\theta, \omega). \tag{2.7}$$

The elements of $\mathbf{d}(\theta, \omega)$ correspond to the complex exponentials $e^{j\omega[\Delta_l(\theta)+p]}$. In general it can be expressed as

$$\mathbf{d}(\theta, \omega) = [1 \; e^{j\omega\tau_2(\theta)} \; e^{j\omega\tau_3(\theta)} \cdots e^{j\omega\tau_N(\theta)}]^{\mathrm{H}} \tag{2.8}$$

where the $\tau_i(\theta)$, $2 \le i \le N$ are the time delays due to propagation and any tap delays from the zero phase reference to the point at which the ith weight is applied. We refer to $\mathbf{d}(\theta, \omega)$ as the array response vector. It is also known as the steering vector, direction vector, or array manifold vector. Nonideal sensor characteristics can be incorporated into $\mathbf{d}(\theta, \omega)$ by multiplying each phase shift by a function $a_i(\theta, \omega)$, which describes the associated sensor response as a function of frequency and direction.

* An FIR filter is by definition linear, so an input sinusoid produces at the output a sinusoid of the same frequency. The magnitude and argument of $r(\omega)$ are, respectively, the magnitude and phase responses.

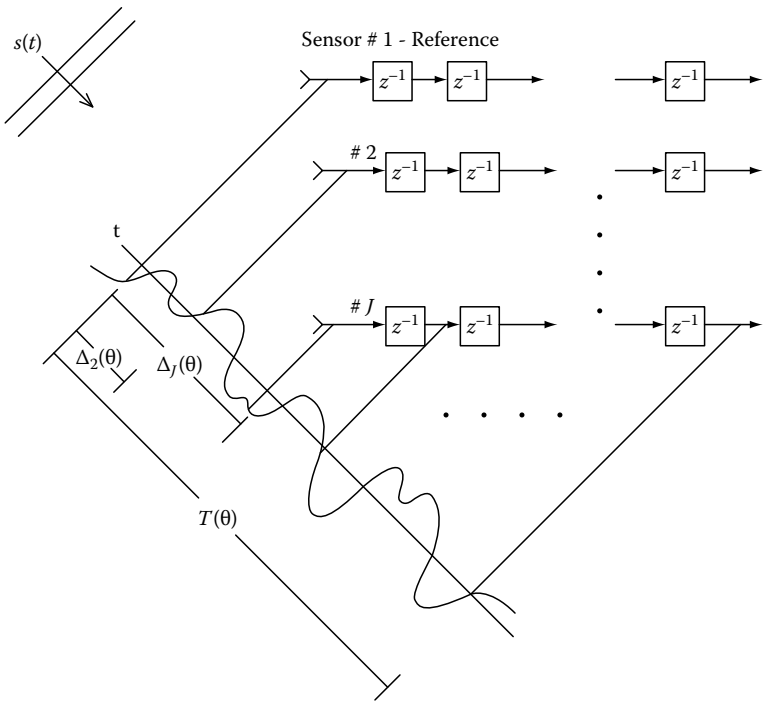

FIGURE 2.2 An array with attached delay lines provides a spatial/temporal sampling of propagating sources. This figure illustrates this sampling of a signal propagating in plane waves from a source located at DOA θ. With J sensors and K samples per sensor, at any instant in time the propagating source signal is sampled at JK nonuniformly spaced points. $T(\theta)$, the time duration from the first sample of the first sensor to the last sample of the last sensor, is termed the temporal aperture of the observation of the source at θ. As notation suggests, temporal aperture will be a function of DOA θ. Plane wave propagation implies that at any time k a propagating signal, received anywhere on a planar front perpendicular to a line drawn from the source to a point on the plane, has equal intensity. Propagation of the signal between two points in space is then characterized as pure delay. In this figure, $\Delta_l(\theta)$ represents the time delay due to plane wave propagation from the first (reference) to the lth sensor.

The "beampattern" is defined as the magnitude squared of $r(\theta, \omega)$. Note that each weight in $\mathbf{w}$ affects both the temporal and spatial responses of the beamformer. Historically, use of FIR filters has been viewed as providing frequency dependent weights in each channel. This interpretation is somewhat incomplete since the coefficients in each filter also influence the spatial filtering characteristics of the beamformer. As a multi-input single output system, the spatial and temporal filtering that occurs is a result of mutual interaction between spatial and temporal sampling.

The correspondence between FIR filtering and beamforming is closest when the beamformer operates at a single temporal frequency ω_o and the array geometry is linear and equispaced as illustrate in Figure 2.3. Letting the sensor spacing be d, propagation velocity be c, and θ represent DOA relative to broadside (perpendicular to the array), we have $\tau_i(\theta) = (i-1)(d/c)\sin\theta$. In this case we identify the relationship between temporal frequency ω in $\mathbf{d}(\omega)$ (FIR filter) and direction θ in $\mathbf{d}(\theta, \omega_o)$ (beamformer) as $\omega = \omega_o(d/c)\sin\theta$. Thus, temporal frequency in an FIR filter corresponds to the sine of direction in a narrowband linear equispaced beamformer. Complete interchange of beamforming and FIR filtering methods is possible for this special case provided the mapping between frequency and direction is accounted for.

The vector notation introduced in Equation 2.3 suggests a vector space interpretation of beamforming. This point of view is useful both in beamformer design and analysis. We use it here in consideration of spatial sampling and array geometry. The weight vector $\mathbf{w}$ and the array response vectors $\mathbf{d}(\theta, \omega)$ are vectors in an

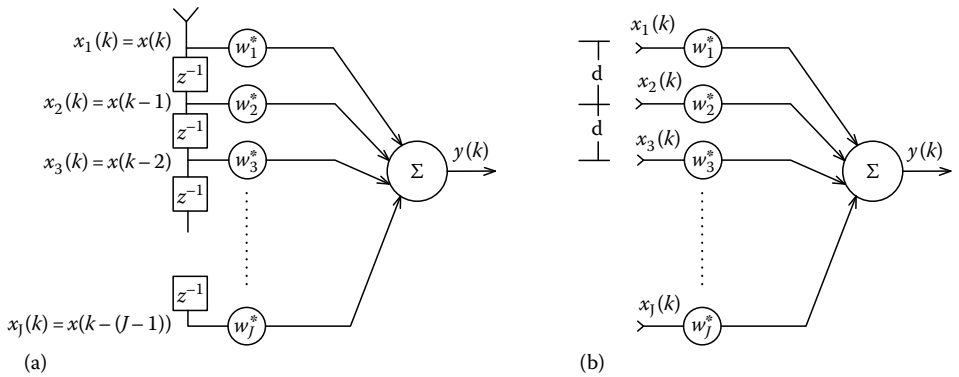

FIGURE 2.3 The analogy between (a) an equispaced omnidirectional narrowband line array and (b) a single-channel FIR filter is illustrated in this figure.

N-dimensional vector space. The angles between **w** and **d**(θ, ω) determine the response $r(\theta, \omega)$. For example, if for some (θ, ω) the angle between **w** and **d**(θ, ω) 90° (i.e., if **w** is orthogonal to **d**(θ, ω)), then the response is zero. If the angle is close to 0°, then the response magnitude will be relatively large. The ability to discriminate between sources at different locations and/or frequencies, say (θ_1, ω_1) and (θ_2, ω_2), is determined by the angle between their array response vectors, $\mathbf{d}(\theta_1, \omega_1)$ and $\mathbf{d}(\theta_2, \omega_2)$.

The general effects of spatial sampling are similar to temporal sampling. Spatial aliasing corresponds to an ambiguity in source locations. The implication is that sources at different locations have the same array response vector, e.g., for narrowband sources $\mathbf{d}(\theta_1, \omega_o)$ and $\mathbf{d}(\theta_2, \omega_o)$. This can occur if the sensors are spaced too far apart. If the sensors are too close together, spatial discrimination suffers as a result of the smaller than necessary aperture; array response vectors are not well dispersed in the *N*-dimensional vector space. Another type of ambiguity occurs with broadband signals when a source at one location and frequency cannot be distinguished from a source at a different location and frequency, i.e., $\mathbf{d}(\theta_1, \omega_1) = \mathbf{d}(\theta_2, \omega_2)$. For example, this occurs in a linear equispaced array whenever $\omega_1 \sin\theta_1 = \omega_2 \sin\theta_2$. (The addition of temporal samples at one sensor prevents this particular ambiguity.)

A primary focus of this section is on designing response via weight selection; however, Equation 2.7 indicates that response is also a function of array geometry (and sensor characteristics if the ideal omnidirectional sensor model is invalid). In contrast with single channel filtering where A/D converters provide a uniform sampling in time, there is no compelling reason to space sensors regularly. Sensor locations provide additional degrees of freedom in designing a desired response and can be selected so that over the range of (θ, ω) of interest the array response vectors are unambiguous and well dispersed in the *N*-dimensional vector space. Utilization of these degrees of freedom can become very complicated due to the multidimensional nature of spatial sampling and the nonlinear relationship between $r(\theta, \omega)$ and sensor locations.

2.2.2 Second-Order Statistics

Evaluation of beamformer performance usually involves power or variance, so the second-order statistics of the data play an important role. We assume the data received at the sensors are zero mean throughout this section. The variance or expected power of the beamformer output is given by $E\{|y|^2\} = \mathbf{w}^H E\{\mathbf{x}\,\mathbf{x}^H\}\mathbf{w}$. If the data are wide sense stationary, then $\mathbf{R}_x = E\{\mathbf{x}\,\mathbf{x}^H\}$, the data covariance matrix, is independent of time. Although we often encounter nonstationary data, the wide sense stationary assumption is used in developing statistically optimal beamformers and in evaluating steady state performance.

Suppose $\mathbf{x}$ represents samples from a uniformly sampled time series having a power spectral density $S(\omega)$ and no energy outside of the spectral band $[\omega_a, \omega_b]$. $\mathbf{R}_x$ can be expressed in terms of the power spectral density of the data using the Fourier transform relationship as

$$\mathbf{R}_x = \frac{1}{2\pi} \int_{\omega_a}^{\omega_b} S(\omega)\ \mathbf{d}(\omega)\ \mathbf{d}^H(\omega) d\omega, \tag{2.9}$$

with $\mathbf{d}(\omega)$ as defined for Equation 2.5. Now assume the array data $\mathbf{x}$ is due to a source located at direction θ. In like manner to the time series case we can obtain the covariance matrix of the array data as

$$\mathbf{R}_x = \frac{1}{2\pi} \int_{\omega_a}^{\omega_b} S(\omega)\ \mathbf{d}(\theta, \omega)\ \mathbf{d}^H(\theta, \omega) d\omega. \tag{2.10}$$

A source is said to be narrowband of frequency ω_o if $\mathbf{R}_x$ can be represented as the rank one outer product

$$\mathbf{R}_x = \sigma_s^2\ \mathbf{d}(\theta, \omega_o)\mathbf{d}^H(\theta, \omega_o), \tag{2.11}$$

where σ_s^2 is the source variance or power.

The conditions under which a source can be considered narrowband depend on both the source bandwidth and the time over which the source is observed. To illustrate this, consider observing an amplitude modulated sinusoid or the output of a narrowband filter driven by white noise on an oscilloscope. If the signal bandwidth is small relative to the center frequency (i.e., if it has small fractional bandwidth), and the time intervals over which the signal is observed are short relative to the inverse of the signal bandwidth, then each observed waveform has the shape of a sinusoid. Note that as the observation time interval is increased, the bandwidth must decrease for the signal to remain sinusoidal in appearance. It turns out, based on statistical arguments, that the observation time bandwidth product (TBWP) is the fundamental parameter that determines whether a source can be viewed as narrowband (see Buckley [2]).

An array provides an effective temporal aperture over which a source is observed. Figure 2.2 illustrates this temporal aperture $T(\theta)$ for a source arriving from direction θ. Clearly the TBWP is dependent on the source DOA. An array is considered narrowband if the observation TBWP is much less than one for all possible source directions.

Narrowband beamforming is conceptually simpler than broadband since one can ignore the temporal frequency variable. This fact, coupled with interest in temporal frequency analysis for some applications, has motivated implementation of broadband beamformers with a narrowband decomposition structure, as illustrated in Figure 2.4. The narrowband decomposition is often performed by taking a discrete Fourier transform (DFT) of the data in each sensor channel using an fast Fourier transform (FFT) algorithm. The data across the array at each frequency of interest are processed by their own beamformer. This is usually termed frequency domain beamforming. The frequency domain beamformer outputs can be made equivalent to the DFT of the broadband beamformer output depicted in Figure 2.1b with proper selection of beamformer weights and careful data partitioning.

2.2.3 Beamformer Classification

Beamformers can be classified as either data independent or statistically optimum, depending on how the weights are chosen. The weights in a data-independent beamformer do not depend on the array data and are chosen to present a specified response for all signal/interference scenarios. The weights in a statistically optimum beamformer are chosen based on the statistics of the array data to "optimize" the array response.

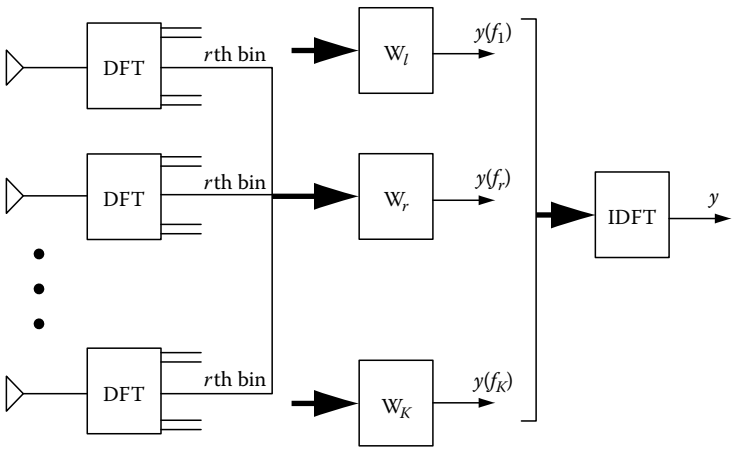

FIGURE 2.4 Beamforming is sometimes performed in the frequency domain when broadband signals are of interest. This figure illustrates transformation of the data at each sensor into the frequency domain. Weighted combinations of data at each frequency (bin) are performed. An inverse discrete Fourier transform produces the output time series.

In general, the statistically optimum beamformer places nulls in the directions of interfering sources in an attempt to maximize the signal-to-noise ratio (SNR) at the beamformer output. A comparison between data-independent and statistically optimum beamformers is illustrated in Figure 2.5.

Sections 2.3 through 2.6 cover data-independent, statistically optimum, adaptive, and partially adaptive beamforming. Data-independent beamformer design techniques are often used in statistically optimum beamforming (e.g., constraint design in linearly constrained minimum variance (LCMV) beamforming). The statistics of the array data are not usually known and may change over time so adaptive algorithms are typically employed to determine the weights. The adaptive algorithm is designed so the beamformer response converges to a statistically optimum solution. Partially adaptive beamformers reduce the adaptive algorithm computational load at the expense of a loss (designed to be small) in statistical optimality.

2.3 Data-Independent Beamforming

The weights in a data-independent beamformer are designed so the beamformer response approximates a desired response independent of the array data or data statistics. This design objective—approximating a desired response—is the same as that for classical finite impulse response (FIR) filter design (see, e.g., Parks and Burrus [8]). We shall exploit the analogies between beamforming and FIR filtering where possible in developing an understanding of the design problem. We also discuss aspects of the design problem specific to beamforming.

The first part of this section discusses forming beams in a classical sense, i.e., approximating a desired response of unity at a point of direction and zero elsewhere. Methods for designing beamformers having more general forms of desired response are presented in the second part.

2.3.1 Classical Beamforming

Consider the problem of separating a single complex frequency component from other frequency components using the J tap FIR filter illustrated in Figure 2.3. If frequency ω_o is of interest, then the desired frequency response is unity at ω_o and zero elsewhere. A common solution to this problem is to choose $\mathbf{w}$ as the vector $\mathbf{d}(\omega_o)$. This choice can be shown to be optimal in terms of minimizing the squared

error between the actual response and desired response. The actual response is characterized by a main lobe (or beam) and many sidelobes. Since $\mathbf{w} = \mathbf{d}(\omega_o)$, each element of $\mathbf{w}$ has unit magnitude. Tapering or windowing the amplitudes of the elements of $\mathbf{w}$ permits trading of main lobe or beam width against sidelobe levels to form the response into a desired shape. Let $\mathbf{T}$ be a J by J diagonal matrix

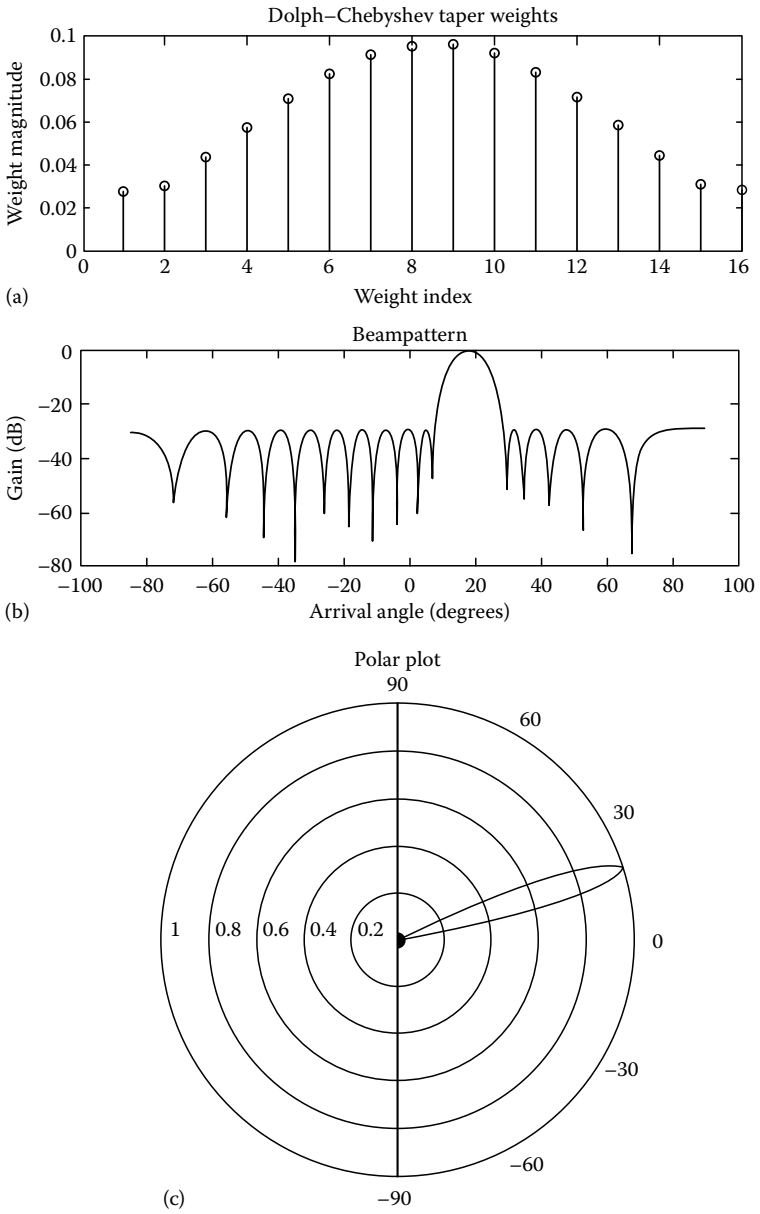

FIGURE 2.5 Beamformers come in both data-independent and statistically optimum varieties. In (a) through (e) we consider an equispaced narrowband array of 16 sensors spaced at one-half wavelength. In (a), (b), and (c) the magnitude of the weights, the beampattern, and the beampattern, in polar coordinates are shown, respectively, for a Dolph–Chebyshev beamformer with −30 dB sidelobes.

(*continued*)

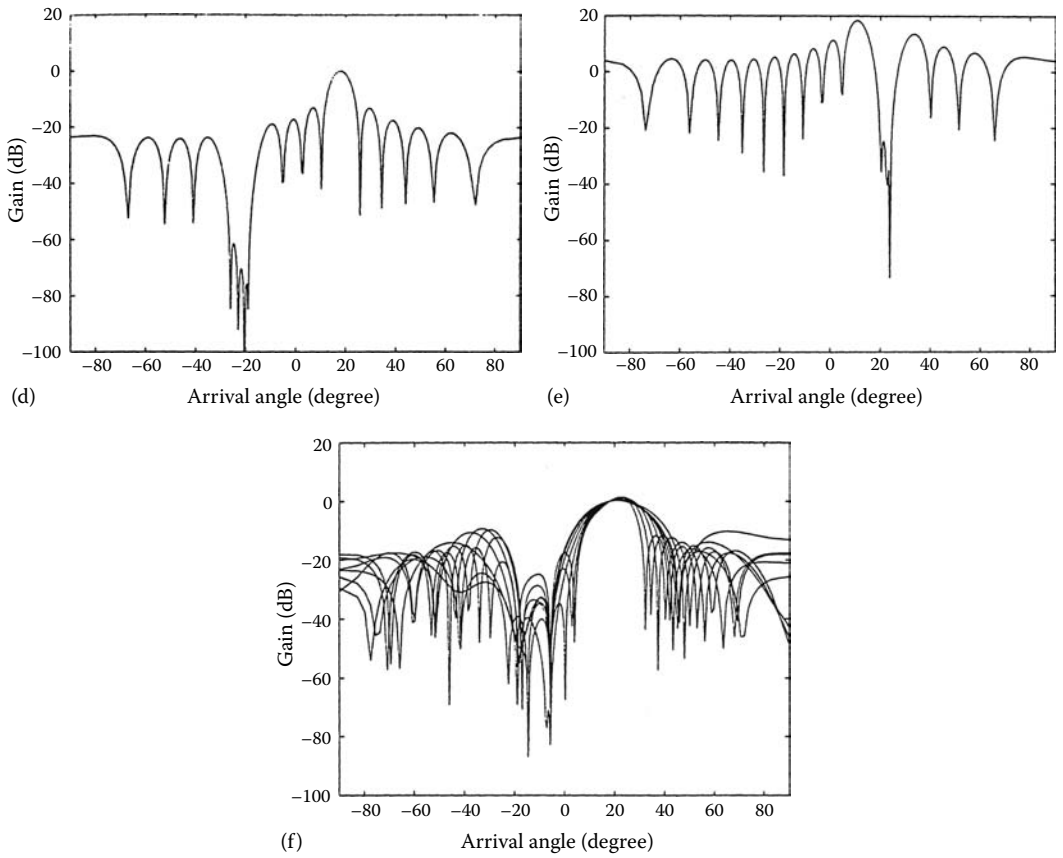

FIGURE 2.5 (continued) In (d) and (e) beampatterns are shown of statistically optimum beamformers which were designed to minimize output power subject to a constraint that the response be unity for an arrival angle of 18°. Energy is assumed to arrive at the array from several interference sources. In (d) several interferers are located between −20° and −23°, each with power of 30 dB relative to the uncorrelated noise power at a single sensor. Deep nulls are formed in the interferer directions. The interferers in (e) are located between 20° and 23°, again with relative power of 30 dB. Again deep nulls are formed at the interferer directions; however, the sidelobe levels are significantly higher at other directions. (f) depicts the broadband LCMV beamformer magnitude response at eight frequencies on the normalized frequency interval $[2\pi/5, 4\pi/5]$ when two interferers arrive from directions −5.75° and −17.5° in the presence of white noise. The interferers have a white spectrum on $[2\pi/5, 4\pi/5]$ and have powers of 40 and 30 dB relative to the white noise, respectively. The constraints are designed to present a unit gain and linear phase over $[2\pi/5, 4\pi/5]$ at a DOA of 18°. The array is linear equispaced with 16 sensors spaced at one-half wavelength for frequency $4\pi/5$ and five tap FIR filters are used in each sensor channel.

with the real-valued taper weights as diagonal elements. The tapered FIR filter weight vector is given by $\mathbf{Td}(\omega_o)$. A detailed comparison of a large number of tapering functions is given in [5].

In spatial filtering one is often interested in receiving a signal arriving from a known location point θ_o. Assuming the signal is narrowband (frequency ω_o), a common choice for the beamformer weight vector is the array response vector $\mathbf{d}(\theta_o, \omega_o)$. The resulting array and beamformer is termed a phased array because the output of each sensor is phase shifted prior to summation. Figure 2.5b depicts the magnitude of the actual response when $\mathbf{w} = \mathbf{Td}(\theta_o, \omega_o)$, where $\mathbf{T}$ implements a common Dolph–Chebyshev tapering function. As in the FIR filter discussed above, beam width and sidelobe levels are the important

characteristics of the response. Amplitude tapering can be used to control the shape of the response, i.e., to form the beam. The equivalence of the narrowband linear equispaced array and FIR filter (see Figure 2.3) implies that the same techniques for choosing taper functions are applicable to either problem. Methods for choosing tapering weights also exist for more general array configurations.

2.3.2 General Data-Independent Response Design

The methods discussed in this section apply to design of beamformers that approximate an arbitrary desired response. This is of interest in several different applications. For example, we may wish to receive any signal arriving from a range of directions, in which case the desired response is unity over the entire range. As another example, we may know that there is a strong source of interference arriving from a certain range of directions, in which case the desired response is zero in this range. These two examples are analogous to bandpass and bandstop FIR filtering. Although we are no longer "forming beams," it is conventional to refer to this type of spatial filter as a beamformer.

Consider choosing $\mathbf{w}$ so the actual response $r(\theta, \omega) = \mathbf{w}^{\mathrm{H}}\mathbf{d}(\theta, \omega)$ approximates desired response $r_{\mathrm{d}}(\theta, \omega)$. Ad hoc techniques similar to those employed in FIR filter design can be used for selecting $\mathbf{w}$. Alternatively, formal optimization design methods can be employed (see, e.g., Parks and Burrus [8]). Here, to illustrate the general optimization design approach, we only consider choosing $\mathbf{w}$ to minimize the weighted averaged square of the difference between desired and actual response.

Consider minimizing the squared error between the actual and desired response at P points (θ_i, ω_i), $1 < i < P$. If $P > N$, then we obtain the overdetermined least squares problem

$$\min_{\mathbf{w}} |\mathbf{A}^{\mathrm{H}}\mathbf{w} - \mathbf{r}_{\mathrm{d}}|^2, \tag{2.12}$$

where

$$\mathbf{A} = [\mathbf{d}(\theta_1, \omega_1),\ \mathbf{d}(\theta_2, \omega_2) \ldots \mathbf{d}(\theta_P, \omega_P)]; \tag{2.13}$$

$$\mathbf{r}_{\mathrm{d}} = [r_{\mathrm{d}}(\theta_1, \omega_1),\ r_{\mathrm{d}}(\theta_2, \omega_2) \ldots r_{\mathrm{d}}(\theta_P, \omega_P)]^{\mathrm{H}}. \tag{2.14}$$

Provided $\mathbf{A}\mathbf{A}^{\mathrm{H}}$ is invertible (i.e., $\mathbf{A}$ is full rank), then the solution to Equation 2.12 is given as

$$\mathbf{w} = \mathbf{A}^{+}\mathbf{r}_{\mathrm{d}}, \tag{2.15}$$

where $\mathbf{A}^{+} = (\mathbf{A}\mathbf{A}^{\mathrm{H}})^{-1}\mathbf{A}$ is the pseudoinverse of $\mathbf{A}$.

A note of caution is in order at this point. The white noise gain of a beamformer is defined as the output power due to unit variance white noise at the sensors. Thus, the norm squared of the weight vector, $\mathbf{w}^{\mathrm{H}}\mathbf{w}$, represents the white noise gain. If the white noise gain is large, then the accuracy by which $\mathbf{w}$ approximates the desired response is a moot point because the beamformer output will have a poor SNR due to white noise contributions. If $\mathbf{A}$ is ill-conditioned, then $\mathbf{w}$ can have a very large norm and still approximate the desired response. The matrix $\mathbf{A}$ is ill-conditioned when the effective numerical dimension of the space spanned by the $\mathbf{d}(\theta_i, \omega_i)$, $1 \le i \le P$, is less than N. For example, if only one source direction is sampled, then the numerical rank of $\mathbf{A}$ is approximately given by the TBWP for that direction. Low rank approximates of $\mathbf{A}$ and $\mathbf{A}^{+}$ should be used whenever the numerical rank is less than N. This ensures that the norm of $\mathbf{w}$ will not be unnecessarily large.

Specific directions and frequencies can be emphasized in Equation 2.12 by selection of the sample points (θ_i, ω_i) and/or unequally weighting of the error at each (θ_i, ω_i). Parks and Burrus [8] discuss this in the context of FIR filtering.

2.4 Statistically Optimum Beamforming

In statistically optimum beamforming, the weights are chosen based on the statistics of the data received at the array. Loosely speaking, the goal is to "optimize" the beamformer response so the output contains minimal contributions due to noise and interfering signals. We discuss several different criteria for choosing statistically optimum beamformer weights. Table 2.1 summarizes these different approaches. Where possible, equations describing the criteria and weights are confined to Table 2.1. Throughout the section we assume that the data is wide-sense stationary and that its second-order statistics are known. Determination of weights when the data statistics are unknown or time varying is discussed in the following section on adaptive algorithms.

2.4.1 Multiple Sidelobe Canceller

The multiple sidelobe canceller (MSC) is perhaps the earliest statistically optimum beamformer. An MSC consists of a "main channel" and one or more "auxiliary channels" as depicted in Figure 2.6a. The main channel can be either a single high gain antenna or a data-independent beamformer (see Section 2.3). It has a highly directional response, which is pointed in the desired signal direction. Interfering signals are assumed to enter through the main channel sidelobes. The auxiliary channels also receive the interfering signals. The goal is to choose the auxiliary channel weights to cancel the main channel interference component. This implies that the responses to interferers of the main channel and linear combination of auxiliary channels must be identical. The overall system then has a response of zero as illustrated in Figure 2.6b. In general, requiring zero response to all interfering signals is either not possible or can result in significant white noise gain. Thus, the weights are usually chosen to trade off interference suppression for white noise gain by minimizing the expected value of the total output power as indicated in Table 2.1.

Choosing the weights to minimize output power can cause cancellation of the desired signal because it also contributes to total output power. In fact, as the desired signal gets stronger it contributes to a larger fraction of the total output power and the percentage cancellation increases. Clearly this is an undesirable effect. The MSC is very effective in applications where the desired signal is very weak (relative to the interference), since the optimum weights will not pay any attention to it, or when the desired signal is known to be absent during certain time periods. The weights can then be adapted in the absence of the desired signal and frozen when it is present.

2.4.2 Use of a Reference Signal

If the desired signal were known, then the weights could be chosen to minimize the error between the beamformer output and the desired signal. Of course, knowledge of the desired signal eliminates the need for beamforming. However, for some applications, enough may be known about the desired signal to generate a signal that closely represents it. This signal is called a reference signal. As indicated in Table 2.1, the weights are chosen to minimize the mean square error between the beamformer output and the reference signal.

The weight vector depends on the cross covariance between the unknown desired signal present in $\mathbf{x}$ and the reference signal. Acceptable performance is obtained provided this approximates the covariance of the unknown desired signal with itself. For example, if the desired signal is amplitude modulated, then acceptable performance is often obtained by setting the reference signal equal to the carrier. It is also assumed that the reference signal is uncorrelated with interfering signals in $\mathbf{x}$. The fact that the direction of the desired signal does not need to be known is a distinguishing feature of the reference signal approach. For this reason it is sometimes termed "blind" beamforming. Other closely related blind beamforming techniques choose weights by exploiting properties of the desired signal such as constant modulus, cyclostationarity, or third and higher order statistics.

TABLE 2.1 Summary of Optimum Beamformers

Type	MSC	Reference Signal	Max SNR	LCMV
Definitions	$\mathbf{x}_a$—auxiliary data	$\mathbf{x}$—array data	$\mathbf{x} = \mathbf{s} + \mathbf{x}$—array data	$\mathbf{x}$—array data
	$\mathbf{y}_m$—primary data	$\mathbf{y}_d$—desired signal	$\mathbf{s}$—signal component	$\mathbf{C}$—constraint matrix
	$\mathbf{r}_{ma} = E\{\mathbf{x}_a \mathbf{y}_m^*\}$	$\mathbf{r}_{xd} = E\{\mathbf{x}\mathbf{y}_d^*\}$	$\mathbf{n}$—noise component	$\mathbf{f}$—response vector
	$\mathbf{R}_a = E\{\mathbf{x}_a \mathbf{x}_a^H\}$	$\mathbf{R}_x = E\{\mathbf{x}\mathbf{x}^H\}$	$\mathbf{R}_s = E\{\mathbf{s}\mathbf{s}^H\}$	$\mathbf{R}_x = E\{\mathbf{x}\mathbf{x}^H\}$
	Output: $\mathbf{y} = \mathbf{y}_m - \mathbf{w}_a^H \mathbf{x}_a$	Output: $\mathbf{y} = \mathbf{w}^H \mathbf{x}$	$\mathbf{R}_n = E\{\mathbf{n}\mathbf{n}^H\}$	Output: $\mathbf{y} = \mathbf{w}^H \mathbf{x}$
			Output: $\mathbf{y} = \mathbf{w}^H \mathbf{x}$	
Criterion	$\min_{\mathbf{w}_a} E\left\{\left\|\mathbf{y}_m - \mathbf{w}_a^H \mathbf{x}_a\right\|^2\right\}$	$\min_{\mathbf{w}} E\left\{\left\|\mathbf{y} - \mathbf{y}_d\right\|^2\right\}$	$\max_{\mathbf{w}} \frac{\mathbf{w}^H \mathbf{R}_s \mathbf{w}}{\mathbf{w}^H \mathbf{R}_n \mathbf{w}}$	$\min_{\mathbf{w}} \{\mathbf{w}^H \mathbf{R}_x \mathbf{w}\}$ s.t. $\mathbf{C}^H \mathbf{w} = \mathbf{f}$
Optimum weights	$\mathbf{w}_a = \mathbf{R}_a^{-1} \mathbf{r}_{ma}$	$\mathbf{w}_a = \mathbf{R}_x^{-1} \mathbf{r}_{rd}$	$\mathbf{R}_n^{-1} \mathbf{R}_s \mathbf{w} = \lambda_{max} \mathbf{w}$	$\mathbf{w} = \mathbf{R}_x^{-1} \mathbf{C} \left[\mathbf{C}^H \mathbf{R}_x^{-1} \mathbf{C}\right]^{-1} f$
Advantages	Simple	Direction of desired signal can be unknown	True maximization of SNR	Flexible and general constraints
Disadvantages	Requires absence of desired signal from auxiliary channels for weight determination	Must generate reference signal	Must know $\mathbf{R}_s$ and $\mathbf{R}_n$ Solve generalized eigenproblem for weights	Computation of constrained weight vector
References	Applebaum (1976)	Widrow (1967)	Monzingo and Miller (1980)	Frost (1972)

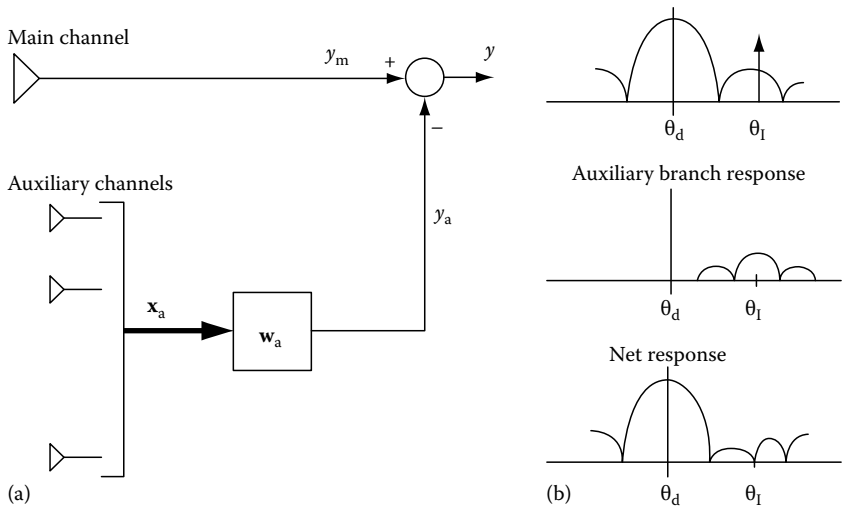

FIGURE 2.6 The multiple sidelobe canceller consists of a main channel and several auxiliary channels as illustrated in (a). The auxiliary channel weights are chosen to "cancel" interference entering through sidelobes of the main channel. (b) Depicts the main channel, auxiliary branch, and overall system response when an interferer arrives from direction θ_I.

2.4.3 Maximization of Signal-to-Noise Ratio

Here the weights are chosen to directly maximize the SNR as indicated in Table 2.1. A general solution for the weights requires knowledge of both the desired signal, $\mathbf{R}_s$, and noise, $\mathbf{R}_n$, covariance matrices. The attainability of this knowledge depends on the application. For example, in an active radar system $\mathbf{R}_n$ can be estimated during the time that no signal is being transmitted and $\mathbf{R}_s$ can be obtained from knowledge of the transmitted pulse and direction of interest. If the signal component is narrowband, of frequency ω, and direction θ, then $\mathbf{R}_s = \sigma^2 \mathbf{d}(\theta, \omega)\mathbf{d}^H(\theta, \omega)$ from the results in Section 2.2. In this case, the weights are obtained as

$$\mathbf{w} = \alpha \mathbf{R}_n^{-1} \mathbf{d}(\theta, \omega), \tag{2.16}$$

where the α is some nonzero complex constant. Substitution of Equation 2.16 into the SNR expression shows that the SNR is independent of the value chosen for α.

2.4.4 Linearly Constrained Minimum Variance Beamforming

In many applications none of the above approaches are satisfactory. The desired signal may be of unknown strength and may always be present, resulting in signal cancellation with the MSC and preventing estimation of signal and noise covariance matrices in the maximum SNR processor. Lack of knowledge about the desired signal may prevent utilization of the reference signal approach. These limitations can be overcome through the application of linear constraints to the weight vector. Use of linear constraints is a very general approach that permits extensive control over the adapted response of the beamformer. In this section we illustrate how linear constraints can be employed to control beamformer response, discuss the optimum linearly constrained beamforming problem, and present the generalized sidelobe canceller (GSC) structure.

The basic idea behind LCMV beamforming is to constrain the response of the beamformer so signals from the direction of interest are passed with specified gain and phase. The weights are chosen to

minimize output variance or power subject to the response constraint. This has the effect of preserving the desired signal while minimizing contributions to the output due to interfering signals and noise arriving from directions other than the direction of interest. The analogous FIR filter has the weights chosen to minimize the filter output power subject to the constraint that the filter response to signals of frequency ω_o be unity.

In Section 2.2, we saw that the beamformer response to a source at angle θ and temporal frequency ω is given by $\mathbf{w}^H\mathbf{d}(\theta,\omega)$. Thus, by linearly constraining the weights to satisfy $\mathbf{w}^H\mathbf{d}(\theta,\omega) = g$ where g is a complex constant, we ensure that any signal from angle θ and frequency ω is passed to the output with response g. Minimization of contributions to the output from interference (signals not arriving from θ with frequency ω) is accomplished by choosing the weights to minimize the output power or variance $E\{|y|^2\} = \mathbf{w}^H\mathbf{R}_x\mathbf{w}$. The LCMV problem for choosing the weights is thus written

$$\min_{\mathbf{w}} \ \mathbf{w}^H\mathbf{R}_x\mathbf{w} \quad \text{subject to} \quad \mathbf{d}^H(\theta,\omega)\mathbf{w} = g^\star. \tag{2.17}$$

The method of Lagrange multipliers can be used to solve Equation 2.17 resulting in

$$\mathbf{w} = g^\star \frac{\mathbf{R}_x^{-1}\mathbf{d}(\theta,\omega)}{\mathbf{d}^H(\theta,\omega)\mathbf{R}_x^{-1}\mathbf{d}(\theta,\omega)}. \tag{2.18}$$

Note that, in practice, the presence of uncorrelated noise will ensure that $\mathbf{R}_x$ is invertible. If $g = 1$, then Equation 2.18 is often termed the minimum variance distortionless response (MVDR) beamformer. It can be shown that Equation 2.18 is equivalent to the maximum SNR solution given in Equation 2.16 by substituting $\sigma^2\mathbf{d}(\theta,\omega)\mathbf{d}^H(\theta,\omega) + \mathbf{R}_n$ for $\mathbf{R}_x$ in Equation 2.18 and applying the matrix inversion lemma.

The single linear constraint in Equation 2.17 is easily generalized to multiple linear constraints for added control over the beampattern. For example, if there is fixed interference source at a known direction ϕ, then it may be desirable to force zero gain in that direction in addition to maintaining the response g to the desired signal. This is expressed as

$$\begin{bmatrix} \mathbf{d}^H(\theta,\omega) \\ \mathbf{d}^H(\phi,\omega) \end{bmatrix} \mathbf{w} = \begin{bmatrix} g^\star \\ 0 \end{bmatrix}. \tag{2.19}$$

If there are $L < N$ linear constraints on $\mathbf{w}$, we write them in the form $\mathbf{C}^H\mathbf{w} = \mathbf{f}$ where the N by L matrix $\mathbf{C}$ and L-dimensional vector $\mathbf{f}$ are termed the constraint matrix and response vector. The constraints are assumed to be linearly independent so $\mathbf{C}$ has rank L. The LCMV problem and solution with this more general constraint equation are given in Table 2.1.

Several different philosophies can be employed for choosing the constraint matrix and response vector. Specifically point, derivative, and eigenvector constraint approaches are popular. Each linear constraint uses one degree of freedom in the weight vector so with L constraints there are only $N - L$ degrees of freedom available for minimizing variance. See Van Veen and Buckley [11] or Van Veen [12] for a more in-depth discussion on this topic.

Generalized sidelobe canceller. The GSC represents an alternative formulation of the LCMV problem, which provides insight, is useful for analysis, and can simplify LCMV beamformer implementation. It also illustrates the relationship between MSC and LCMV beamforming. Essentially, the GSC is a mechanism for changing a constrained minimization problem into unconstrained form.

Suppose we decompose the weight vector $\mathbf{w}$ into two orthogonal components $\mathbf{w}_o$ and $-\mathbf{v}$ (i.e., $\mathbf{w} = \mathbf{w}_o - \mathbf{v}$) that lie in the range and null spaces of $\mathbf{C}$, respectively. The range and null spaces of a matrix span the entire space so this decomposition can be used to represent any $\mathbf{w}$. Since $\mathbf{C}^H\mathbf{v} = 0$, we must have

$$\mathbf{w}_o = \mathbf{C}(\mathbf{C}^H\mathbf{C})^{-1}\mathbf{f}, \tag{2.20}$$

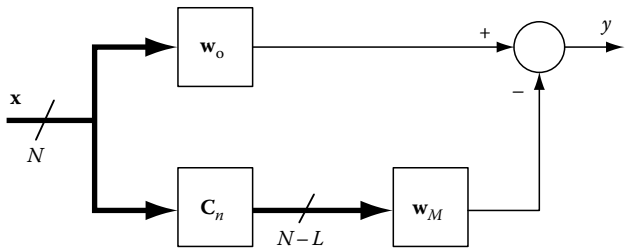

FIGURE 2.7 The generalized sidelobe canceller represents an implementation of the LCMV beamformer in which the adaptive weights are unconstrained. It consists of a preprocessor composed of a fixed beamformer $\mathbf{w}_o$ and a blocking matrix $\mathbf{C}_n$, and a standard adaptive filter with unconstrained weight vector $\mathbf{w}_M$.

if $\mathbf{w}$ is to satisfy the constraints. Equation 2.20 is the minimum L_2 norm solution to the underdetermined equivalent of Equation 2.12. The vector $\mathbf{v}$ is a linear combination of the columns of an N by M ($M = N - L$) matrix $\mathbf{C}_n$ (i.e., $\mathbf{v} = \mathbf{C}_n\mathbf{w}_M$) provided the columns of $\mathbf{C}_n$ form a basis for the null space of $\mathbf{C}$. $\mathbf{C}_n$ can be obtained from $\mathbf{C}$ using any of several orthogonalization procedures such as Gram–Schmidt, QR decomposition, or singular value decomposition. The weight vector $\mathbf{w} = \mathbf{w}_o - \mathbf{C}_n\mathbf{w}_M$ is depicted in block diagram form in Figure 2.7. The choice for $\mathbf{w}_o$ and $\mathbf{C}_n$ implies that $\mathbf{w}$ satisfies the constraints independent of $\mathbf{w}_M$ and reduces the LCMV problem to the unconstrained problem

$$\min_{\mathbf{w}_M} [\mathbf{w}_o - \mathbf{C}_n\mathbf{w}_M]^H \mathbf{R}_x [\mathbf{w}_o - \mathbf{C}_n\mathbf{w}_M]. \tag{2.21}$$

The solution is

$$\mathbf{w}_M = \left(\mathbf{C}_n^H \mathbf{R}_x \mathbf{C}_n\right)^{-1} \mathbf{C}_n^H \mathbf{R}_x \mathbf{w}_o. \tag{2.22}$$

The primary implementation advantages of this alternate but equivalent formulation stem from the facts that the weights $\mathbf{w}_M$ are unconstrained and a data-independent beamformer $\mathbf{w}_o$ is implemented as an integral part of the optimum beamformer. The unconstrained nature of the adaptive weights permits much simpler adaptive algorithms to be employed and the data-independent beamformer is useful in situations where adaptive signal cancellation occurs (see Section 2.4.5).

As an example, assume the constraints are as given in Equation 2.17. Equation 2.20 implies $\mathbf{w}_o = g^* \mathbf{d}(\theta, \omega) / \left[\mathbf{d}^H(\theta, \omega)\mathbf{d}(\theta, \omega)\right]$. $\mathbf{C}_n$ satisfies $\mathbf{d}^H(\theta, \omega)\mathbf{C}_n = 0$ so each column $[\mathbf{C}_n]_i$; $1 < i < N - L$, can be viewed as a data-independent beamformer with a null in direction θ at frequency ω: $\mathbf{d}^H(\theta, \omega)[\mathbf{C}_n]_j = 0$. Thus, a signal of frequency ω and direction θ arriving at the array will be blocked or nulled by the matrix $\mathbf{C}_n$. In general, if the constraints are designed to present a specified response to signals from a set of directions and frequencies, then the columns of $\mathbf{C}_n$ will block those directions and frequencies. This characteristic has led to the term "blocking matrix" for describing $\mathbf{C}_n$. These signals are only processed by $\mathbf{w}_o$ and since $\mathbf{w}_o$ satisfies the constraints, they are presented with the desired response independent of $\mathbf{w}_M$. Signals from directions and frequencies over which the response is not constrained will pass through the upper branch in Figure 2.7 with some response determined by $\mathbf{w}_o$. The lower branch chooses $\mathbf{w}_M$ to estimate the signals at the output of $\mathbf{w}_o$ as a linear combination of the data at the output of the blocking matrix. This is similar to the operation of the MSC, in which weights are applied to the output of auxiliary sensors in order to estimate the primary channel output (see Figure 2.6).

2.4.5 Signal Cancellation in Statistically Optimum Beamforming

Optimum beamforming requires some knowledge of the desired signal characteristics, either its statistics (for maximum SNR or reference signal methods), its direction (for the MSC), or its response

vector $\mathbf{d}(\theta, \omega)$ (for the LCMV beamformer). If the required knowledge is inaccurate, the optimum beamformer will attenuate the desired signal as if it were interference. Cancellation of the desired signal is often significant, especially if the SNR of the desired signal is large. Several approaches have been suggested to reduce this degradation (e.g., Cox et al. [3]).

A second cause of signal cancellation is correlation between the desired signal and one or more interference signals. This can result either from multipath propagation of a desired signal or from smart (correlated) jamming. When interference and desired signals are uncorrelated, the beamformer attenuates interferers to minimize output power. However, with a correlated interferer the beamformer minimizes output power by processing the interfering signal in such a way as to cancel the desired signal. If the interferer is partially correlated with the desired signal, then the beamformer will cancel the portion of the desired signal that is correlated with the interferer. Methods for reducing signal cancellation due to correlated interference have been suggested (e.g., Widrow et al. [13], Shan and Kailath [10]).

2.5 Adaptive Algorithms for Beamforming

The optimum beamformer weight vector equations listed in Table 2.1 require knowledge of second-order statistics. These statistics are usually not known, but with the assumption of ergodicity, they (and therefore the optimum weights) can be estimated from available data. Statistics may also change over time, e.g., due to moving interferers. To solve these problems, weights are typically determined by adaptive algorithms.

There are two basic adaptive approaches: (1) block adaptation, where statistics are estimated from a temporal block of array data and used in an optimum weight equation; and (2) continuous adaptation, where the weights are adjusted as the data is sampled such that the resulting weight vector sequence converges to the optimum solution. If a nonstationary environment is anticipated, block adaptation can be used, provided that the weights are recomputed periodically. Continuous adaptation is usually preferred when statistics are time-varying or, for computational reasons, when the number of adaptive weights M is moderate to large; values of $M > 50$ are common.

Among notable adaptive algorithms proposed for beamforming are the Howells–Applebaum adaptive loop developed in the late 1950s and reported by Howells [7] and Applebaum [1], and the Frost LCMV algorithm [4]. Rather than recapitulating adaptive algorithms for each optimum beamformer listed in Table 2.1, we take a unifying approach using the standard adaptive filter configuration illustrated on the right side of Figure 2.7.

In Figure 2.7 the weight vector $\mathbf{w}_M$ is chosen to estimate the desired signal y_d as linear combination of the elements of the data vector $\mathbf{u}$. We select $\mathbf{w}_M$ to minimize the mean squared error (MSE)

$$J(\mathbf{w}_M) = E\left\{\left|y_\mathrm{d} - \mathbf{w}_M^\mathrm{H}\mathbf{u}\right|^2\right\} = \sigma_\mathrm{d}^2 - \mathbf{w}_M^\mathrm{H}\mathbf{r}_\mathrm{ud} - \mathbf{r}_\mathrm{ud}^\mathrm{H}\mathbf{w}_M + \mathbf{w}_M^\mathrm{H}\mathbf{R}_\mathrm{u}\mathbf{w}_M, \tag{2.23}$$

where

$\sigma_\mathrm{d}^2 = E\{|y_\mathrm{d}|^2\}$
$\mathbf{r}_\mathrm{ud} = E\{\mathbf{u}\ y_\mathrm{d}^*\}$
$\mathbf{R}_\mathrm{u} = E\{\mathbf{u}\ \mathbf{u}^\mathrm{H}\}$

$J(\mathbf{w}_M)$ is minimized by

$$\mathbf{w}_\mathrm{opt} = \mathbf{R}_\mathrm{u}^{-1}\mathbf{r}_\mathrm{ud}. \tag{2.24}$$

Comparison of Equation 2.23 and the criteria listed in Table 2.1 indicates that this standard adaptive filter problem is equivalent to both the MSC beamformer problem (with $y_\mathrm{d} = y_m$ and $\mathbf{u} = \mathbf{x}_\mathrm{a}$) and the reference signal beamformer problem (with $\mathbf{u} = \mathbf{x}$). The LCMV problem is apparently different. However closer examination of Figure 2.7 and Equations 2.22 and 2.24 reveals that the standard adaptive filter

problem is equivalent to the LCMV problem implemented with the GSC structure. Setting $\mathbf{u} = \mathbf{C}_n^{\mathrm{H}}\mathbf{x}$ and $y_{\mathrm{d}} = \mathbf{w}_{\mathrm{o}}^{\mathrm{H}}\mathbf{x}$ implies $\mathbf{R}_{\mathrm{u}} = \mathbf{C}_n^{\mathrm{H}}\mathbf{R}_x\mathbf{C}_n$ and $\mathbf{r}_{\mathrm{ud}} = \mathbf{C}_n^{\mathrm{H}}\mathbf{R}_x\mathbf{w}_{\mathrm{o}}$. The maximum SNR beamformer cannot in general be represented by Figure 2.7 and Equation 2.24. However, it was noted after Equation 2.18 that if the desired signal is narrowband, then the maximum SNR and the LCMV beamformers are equivalent.

The block adaptation approach solves Equation 2.24 using estimates of $\mathbf{R}_{\mathrm{u}}$ and $\mathbf{r}_{\mathrm{ud}}$ formed from K samples of $\mathbf{u}$ and y_{d}: $\mathbf{u}(k)$, $y_{\mathrm{d}}(k)$; $0 < k < K-1$. The most common are the sample covariance matrix

$$\hat{\mathbf{R}}_{\mathrm{u}} = \frac{1}{K}\sum_{k=0}^{K-1}\mathbf{u}(k)\mathbf{u}^{\mathrm{H}}(k), \tag{2.25}$$

and sample cross-covariance vector

$$\hat{\mathbf{r}}_{\mathrm{ud}} = \frac{1}{K}\sum_{k=0}^{K-1}\mathbf{u}(k)y_{\mathrm{d}}^{\star}(k). \tag{2.26}$$

Performance analysis and guidelines for selecting the block size K are provided in Reed et al. [9].

Continuous adaptation algorithms are easily developed in terms of Figure 2.7 and Equation 2.23. Note that $J(\mathbf{w}_M)$ is a quadratic error surface. Since the quadratic surface's "Hessian" $\mathbf{R}_{\mathrm{u}}$ is the covariance matrix of noisy data, it is positive definite. This implies that the error surface is a "bowl." The shape of the bowl is determined by the eigenstructure of $\mathbf{R}_{\mathrm{u}}$. The optimum weight vector $\mathbf{w}_{\mathrm{opt}}$ corresponds to the bottom of the bowl.

One approach to adaptive filtering is to envision a point on the error surface that corresponds to the present weight vector $\mathbf{w}_M(k)$. We select a new weight vector $\mathbf{w}_M(k+1)$ so as to descend on the error surface. The gradient vector

$$\nabla_{\mathrm{w}_M(k)} = \frac{\partial}{\partial \mathbf{w}_M}J(\mathbf{w}_M)\bigg|_{\mathrm{w}_M=\mathbf{w}_m(k)} = -2\mathbf{r}_{\mathrm{ud}} + 2\mathbf{R}_{\mathrm{u}}\mathbf{w}_M(k), \tag{2.27}$$

tells us the direction in which to adjust the weight vector. Steepest descent, i.e., adjustment in the negative gradient direction, leads to the popular least mean-square (LMS) adaptive algorithm. The LMS algorithm replaces $\nabla_{\mathrm{w}_M(k)}$ with the instantaneous gradient estimate $\hat{\nabla}_{\mathrm{w}_M(k)} = -2[\mathbf{u}(k)y_{\mathrm{d}}^{\star}(k) - \mathbf{u}(k)\mathbf{u}^{\mathrm{H}}(k)\mathbf{w}_M(k)]$. Denoting $y(k) = y_{\mathrm{d}}(k) - \mathbf{w}_M^{\mathrm{H}}\mathbf{u}(k)$, we have

$$\mathbf{w}_M(k+1) = \mathbf{w}_M(k) + \mu\mathbf{u}(k)y^{\star}(k). \tag{2.28}$$

The gain constant μ controls convergence characteristics of the random vector sequence $\mathbf{w}_M(k)$. Table 2.2 provides guidelines for its selection.

The primary virtue of the LMS algorithm is its simplicity. Its performance is acceptable in many applications; however, its convergence characteristics depend on the shape of the error surface and therefore the eigenstructure of $\mathbf{R}_{\mathrm{u}}$. When the eigenvalues are widely spread, convergence can be slow and other adaptive algorithms with better convergence characteristics should be considered. Alternative procedures for searching the error surface have been proposed in addition to algorithms based on least squares and Kalman filtering. Roughly speaking, these algorithms trade off computational requirements with speed of convergence to $\mathbf{w}_{\mathrm{opt}}$. We refer you to texts on adaptive filtering for detailed descriptions and analysis (Widrow and Stearns [14], Haykin [6], and others).

One alternative to LMS is the exponentially weighted recursive least squares (RLS) algorithm. At the Kth time step, $\mathbf{w}_M(K)$ is chosen to minimize a weighted sum of past squared errors

$$\min_{\mathbf{w}_M(K)}\sum_{k=0}^{K}\lambda^{K-k}\left|y_{\mathrm{d}}(k) - \mathbf{w}_M^{\mathrm{H}}(K)\mathbf{u}(k)\right|^2. \tag{2.29}$$

TABLE 2.2 Comparison of the LMS and RLS Weight Adaptation Algorithms

Algorithm	LMS	RLS
Initialization	$\mathbf{w}_M(0) = 0$	$\mathbf{w}_M(0) = 0$
	$\mathbf{y}(0) = \mathbf{y}_d(0)$	$\mathbf{P}(0) = \delta^{-1}\mathbf{I}$
	$0 < \mu < \frac{1}{\text{Trace}[\mathbf{R}_u]}$	δ small, $\mathbf{I}$ identity matrix
Update Equations	$\mathbf{w}_M(k) = \mathbf{w}_M(k-1) + \mu\mathbf{u}(k-1)y^*(k-1)$	$\mathbf{v}(k) = \mathbf{P}(k-1)\mathbf{u}(k)$
	$y(k) = y_d(k) - \mathbf{w}_M^H(k)\mathbf{u}(k)$	$\mathbf{k}(k) = \frac{\lambda^{-1}\mathbf{v}(k)}{1+\lambda^{-1}\mathbf{u}^H(k)\mathbf{v}(k)}$
		$\alpha(k) = y_d(k) - \mathbf{w}_M^H(k-1)\mathbf{u}(k)$
		$\mathbf{w}_M(k) = \mathbf{w}_M(k-1) + \mathbf{k}(k)\alpha^*(k)$
		$\mathbf{P}(k) = \lambda^{-1}\mathbf{P}(k-1) - \lambda^{-1}\mathbf{k}(k)\mathbf{v}^H(k)$
Multiplies per update	$2M$	$4M^2 + 4M + 2$
Performance Characteristics	Under certain conditions, convergence of $\mathbf{w}_M(k)$ to the statistically optimum weight vector $\mathbf{w}_{opt}$ in the mean-square sense is guaranteed if μ is chosen as indicated above. The convergence rate is governed by the eigenvalue spread of $\mathbf{R}_u$. For large eigenvalue spread, convergence can be very slow.	The $\mathbf{w}_M(k)$ represents the least squares solution at each instant k and are optimum in a deterministic sense. Convergence to the statistically optimum weight vector $\mathbf{w}_{opt}$ is often faster than that obtained using the LMS algorithm because it is independent of the eigenvalue spread of $\mathbf{R}_u$.

λ is a positive constant less than one which determines how quickly previous data are de-emphasized. The RLS algorithm is obtained from Equation 2.29 by expanding the magnitude squared and applying the matrix inversion lemma. Table 2.2 summarizes both the LMS and RLS algorithms.

2.6 Interference Cancellation and Partially Adaptive Beamforming

The computational requirements of each update in adaptive algorithms are proportional to either the weight vector dimension M (e.g., LMS) or dimension squared M^2 (e.g., RLS). If M is large, this requirement is quite severe and for practical real time implementation it is often necessary to reduce M. Furthermore, the rate at which an adaptive algorithm converges to the optimum solution may be very slow for large M. Adaptive algorithm convergence properties can be improved by reducing M.

The concept of "degrees of freedom" is much more relevant to this discussion than the number of weights. The expression degrees of freedom refers to the number of unconstrained or "free" weights in an implementation. For example, an LCMV beamformer with L constraints on N weights has $N - L$ degrees of freedom; the GSC implementation separates these as the unconstrained weight vector $\mathbf{w}_M$. There are M degrees of freedom in the structure of Figure 2.7. A fully adaptive beamformer uses all available degrees of freedom and a partially adaptive beamformer uses a reduced set of degrees of freedom. Reducing degrees of freedom lowers computational requirements and often improves adaptive response time. However, there is a performance penalty associated with reducing degrees of freedom. A partially adaptive beamformer cannot generally converge to the same optimum solution as the fully adaptive beamformer. The goal of partially adaptive beamformer design is to reduce degrees of freedom without significant degradation in performance.

The discussion in this section is general, applying to different types of beamformers although we borrow much of the notation from the GSC. We assume the beamformer is described by the adaptive structure of Figure 2.7 where the desired signal y_d is obtained as $y_d = \mathbf{w}_o^H\mathbf{x}$ and the data vector $\mathbf{u}$ as $\mathbf{u} = \mathbf{T}^H\mathbf{x}$. Thus, the beamformer output is $y = \mathbf{w}^H\mathbf{x}$ where $\mathbf{w} = \mathbf{w}_o - \mathbf{T}\mathbf{w}_M$. In order to distinguish between fully and partially adaptive implementations, we decompose $\mathbf{T}$ into a product of two matrices $\mathbf{C}_n\mathbf{T}_M$.

The definition of $\mathbf{C}_n$ depends on the particular beamformer and $\mathbf{T}_M$ represents the mapping which reduces degrees of freedom. The MSC and GSC are obtained as special cases of this representation. In the MSC $\mathbf{w}_o$ is an N vector that selects the primary sensor, $\mathbf{C}_n$ is an N by $N-1$ matrix that selects the $N-1$ possible auxiliary sensors from the complete set of N sensors, and $\mathbf{T}_M$ is an $N-1$ by M matrix that selects the M auxiliary sensors actually utilized. In terms of the GSC, $\mathbf{w}_o$ and $\mathbf{C}_n$ are defined as in Section 2.4.4 and $\mathbf{T}_M$ is an $N-L$ by M matrix that reduces degrees of freedom ($M < N-L$).

The goal of partially adaptive beamformer design is to choose $\mathbf{T}_M$ (or $\mathbf{T}$) such that good interference cancellation properties are retained even though M is small. To see that this is possible in principle, consider the problem of simultaneously canceling two narrowband sources from direction θ_1 and θ_2 at frequency ω_o. Perfect cancellation of these sources requires $\mathbf{w}^H\mathbf{d}(\theta_1, \omega_o) = 0$ and $\mathbf{w}^H\mathbf{d}(\theta_2, \omega_o) = 0$ so we must choose $\mathbf{w}_M$ to satisfy

$$\mathbf{w}_M^H\left[\mathbf{T}^H\mathbf{d}(\theta_1, \omega_o)\mathbf{T}^H\mathbf{d}(\theta_2, \omega_o)\right] = [g_1, g_2], \tag{2.30}$$

where $g_i = \mathbf{w}_o^H\mathbf{d}(\theta_i, \omega_o)$ is the response of the $\mathbf{w}_o$ branch to the ith interferer. Assuming $\mathbf{T}^H\mathbf{d}(\theta_1, \omega_o)$ and $\mathbf{T}^H\mathbf{d}(\theta_2, \omega_o)$ are linearly independent and nonzero, and provided $M \geq 2$, then at least one $\mathbf{w}_M$ exists that satisfies Equation 2.30. Extending this reasoning, we see that $\mathbf{w}_M$ can be chosen to cancel M narrowband interferers (assuming the $\mathbf{T}^H\mathbf{d}(\theta_i, \omega_o)$ are linearly independent and nonzero), independent of $\mathbf{T}$. Total cancellation occurs if $\mathbf{w}_M$ is chosen so the response of $\mathbf{T}\,\mathbf{w}_M$ perfectly matches the $\mathbf{w}_o$ branch response to the interferers. In general, M narrowband interferers can be canceled using M adaptive degrees of freedom with relatively mild restrictions on $\mathbf{T}$.

No such rule exists in the broadband case. Here complete cancellation of a single interferer requires choosing $\mathbf{T}\mathbf{w}_M$ so that the response of the adaptive branch, $\mathbf{w}_M^H\mathbf{T}^H\mathbf{d}(\theta_1, \omega)$, matches the response of the $\mathbf{w}_o$ branch, $\mathbf{w}_o^H\mathbf{d}(\theta_1, \omega)$, over the entire frequency band of the interferer. In this case, the degree of cancellation depends on how well these two responses match and is critically dependent on the interferer direction, frequency content, and $\mathbf{T}$. Good cancellation can be obtained in some situations when $M = 1$, while in others even large values of M result in poor cancellation.

A variety of intuitive and optimization-based techniques have been proposed for designing $\mathbf{T}_M$ that achieve good interference cancellation with relatively small degrees of freedom. See Van Veen and Buckley [11] and Van Veen [12] for further review and discussion.

2.7 Summary

A beamformer forms a scalar output signal as a weighted combination of the data received at an array of sensors. The weights determine the spatial filtering characteristics of the beamformer and enable separation of signals having overlapping frequency content if they originate from different locations. The weights in a data-independent beamformer are chosen to provide a fixed response independent to the received data. Statistically optimum beamformers select the weights to optimize the beamformer response based on the statistics of the data. The data statistics are often unknown and may change with time so adaptive algorithms are used to obtain weights that converge to the statistically optimum solution. Computational and response time considerations dictate the use of partially adaptive beamformers with arrays composed of large numbers of sensors.

Defining Terms

Array response vector: Vector describing the amplitude and phase relationships between propagating wave components at each sensor as a function of spatial direction and temporal frequency. Forms the basis for determining the beamformer response.

Beamformer: A device used in conjunction with an array of sensors to separate signals and interference on the basis of their spatial characteristics. The beamformer output is usually given by a weighted combination of the sensor outputs.

Beampattern: The magnitude squared of the beamformer's spatial filtering response as a function of spatial direction and possibly temporal frequency.

Data-independent, statistically optimum, adaptive, and partially adaptive beamformers: The weights in a data-independent beamformer are chosen independent of the statistics of the data. A statistically optimum beamformer chooses its weights to optimize some statistical function of the beamformer output, such as SNR. An adaptive beamformer adjusts its weights in response to the data to accommodate unknown or time varying statistics. A partially adaptive beamformer uses only a subset of the available adaptive degrees of freedom to reduce the computational burden or improve the adaptive convergence rate.

Generalized sidelobe canceller: Structure for implementing LCMV beamformers that separates the constrained and unconstrained components of the adaptive weight vector. The unconstrained components adaptively cancel interference that leaks through the sidelobes of a data-independent beamformer designed to satisfy the constraints.

LCMV beamformer: Beamformer in which the weights are chosen to minimize the output power subject to a linear response constraint. The constraint preserves the signal of interest while power minimization optimally attenuates noise and interference.

Multiple sidelobe canceller: Adaptive beamformer structure in which the data received at low gain auxiliary sensors is used to adaptively cancel the interference arriving in the mainlobe or sidelobes of a spatially high gain sensor.

MVDR beamformer: A form of LCMV beamformer employing a single constraint designed to pass a signal of given direction and frequency with unit gain.

References

1. Applebaum, S.P., Adaptive arrays, Syracuse University Research Corp., Report SURC SPL TR 66-001, August 1966 (reprinted in *IEEE Trans. AP,* AP-24, 585–598, September 1976).
2. Buckley, K.M., Spatial/spectral filtering with linearly-constrained minimum variance beamformers, *IEEE Trans. ASSP*, ASSP-35, 249–266, March 1987.
3. Cox, H., Zeskind, R.M., and Owen, M.M., Robust adaptive beamforming, *IEEE Trans. ASSP*, ASSP-35, 1365–1375, October 1987.
4. Frost III, O.L., An algorithm for linearly constrained adaptive array processing, *Proc. IEEE*, 60, 926–935, August 1972.
5. Harris, F.J., On the use of windows for harmonic analysis with the discrete Fourier transform, *Proc. IEEE*, 66, 51–83, January 1978.
6. Haykin, S., *Adaptive Filter Theory*, 3rd edn., Prentice-Hall, Englewood Cliffs, NJ, 1996.
7. Howells, P.W., Explorations in fixed and adaptive resolution at GE and SURC, *IEEE Trans. AP*, AP-24, 575–584, September 1976.
8. Parks, T.W. and Burrus, C.S., *Digital Filter Design*, Wiley-Interscience, New York, 1987.
9. Reed, I.S., Mallett, J.D., and Brennen, L.E., Rapid convergence rate in adaptive arrays, *IEEE Trans. AES*, AES-10, 853–863, November 1974.
10. Shan, T. and Kailath, T., Adaptive beamforming for coherent signals and interference, *IEEE Trans. ASSP*, ASSP-33, 527–536, June 1985.
11. Van Veen, B. and Buckley, K., Beamforming: a versatile approach to spatial filtering, *IEEE ASSP Mag.*, 5(2), 4–24, April 1988.
12. Van Veen, B., Minimum variance beamforming, in *Adaptive Radar Detection and Estimation*, Haykin, S. and Steinhardt, A., eds., John Wiley & Sons, New York, Chap. 4, pp. 161–236, 1992.

13. Widrow, B., Duvall, K.M., Gooch, R.P., and Newman, W.C., Signal cancellation phenomena in adaptive arrays: Causes and cures, *IEEE Trans. AP*, AP-30, 469–478, May 1982.
14. Widrow, B. and Stearns, S., *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

Further Readings

For further information, we refer the reader to the following books

Compton, R.T., Jr. *Adaptive Antennas: Concepts and Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

Haykin, S., ed. *Array Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

Johnson, D. and Dudgeon, D., *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

Monzingo, R. and Miller, T., *Introduction to Adaptive Arrays*, John Wiley & Sons, New York, 1980.

Widrow, P.E., Mantey, P.E., Griffiths, L.J., and Goode, B.B., Adaptive Antenna Systems, *Proc. IEEE*, 55(12), 2143–2159, December 1967.

Tutorial Articles

Gabriel, W.F., Adaptive arrays: An introduction, *Proc. IEEE*, 64, 239–272, August 1976 and bibliography.

Marr, J., A selected bibliography on adaptive antenna arrays, *IEEE Trans. AES*, AES-22, 781–798, November 1986.

Several special journal issues have been devoted to beamforming—*IEEE Transactions on Antennas and Propagation,* September 1976 and March 1986, and the *Journal of Ocean Engineering,* 1987. Papers devoted to beamforming are often found in the *IEEE Transactions on: Antennas and Propagation, Signal Processing, Aerospace and Electronic Systems,* and in the *Journal of the Acoustical Society of America*.

3

Subspace-Based Direction-Finding Methods

Egemen Gönen
Globalstar

Jerry M. Mendel
University of Southern California

3.1 Introduction

Estimating bearings of multiple narrowband signals from measurements collected by an array of sensors has been a very active research problem for the last two decades. Typical applications of this problem are radar, communication, and underwater acoustics. Many algorithms have been proposed to solve the bearing estimation problem. One of the first techniques that appeared was beamforming which has a resolution limited by the array structure. Spectral estimation techniques were also applied to the problem. However, these techniques fail to resolve closely spaced arrival angles for low signal-to-noise ratios (SNRs). Another approach is the maximum-likelihood (ML) solution. This approach has been well documented in the literature. In the stochastic ML method [29], the sbgv signals are assumed to be Gaussian whereas they are regarded as arbitrary and deterministic in the deterministic ML method [37]. The sensor noise is modeled as Gaussian in both methods, which is a reasonable assumption due to the central limit theorem. The stochastic ML estimates of the bearings achieve the Cramer–Rao bound (CRB). On the other hand, this does not hold for deterministic ML estimates [32]. The common problem with the ML methods in general is the necessity of solving a nonlinear multidimensional (MD) optimization problem which has a high computational cost and for which there is no guarantee of global convergence. "Subspace-based" (or, super-resolution) approaches have attracted much attention, after the work of Schmidt [29], due to their computational simplicity as compared to the ML approach, and their possibility of overcoming the Rayleigh bound on the resolution power of classical direction-finding methods. Subspace-based direction-finding methods are summarized in this section.

3.2 Formulation of the Problem

Consider an array of M antenna elements receiving a set of plane waves emitted by $P(P < M)$ sources in the far field of the array. We assume a narrowband propagation model, i.e., the signal envelopes do not change during the time it takes for the wave fronts to travel from one sensor to another. Suppose that the signals have a common frequency of f_0; then, the wavelength $\lambda = c/f_0$ where c is the speed of propagation. The received M-vector $\mathbf{r}(t)$ at time t is

$$\mathbf{r}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t), \tag{3.1}$$

where

$\mathbf{s}(t) = [s_1(t), \ldots, s_P(t)]^{\mathrm{T}}$ is the P-vector of sources

$\mathbf{A} = [\mathbf{a}(\theta_1), \ldots, \mathbf{a}(\theta_P)]$ is the $M \times P$ steering matrix in which $\mathbf{a}(\theta_i)$, the ith steering vector, is the response of the array to the ith source arriving from θ_i

$\mathbf{n}(t) = [n_1(t), \ldots, n_M(t)]^{\mathrm{T}}$ is an additive noise process

We assume (1) the source signals may be statistically independent, partially correlated, or completely correlated (i.e., coherent); the distributions are unknown; (2) the array may have an arbitrary shape and response; and (3) the noise process is independent of the sources, zero-mean, and it may be either partially white or colored; its distribution is unknown. These assumptions will be relaxed, as required by specific methods, as we proceed.

The direction finding problem is to estimate the bearings [i.e., directions of arrival (DOA)] $\{\theta_i\}_{i=1}^{P}$ of the sources from the snapshots $\mathbf{r}(t)$, $t = 1, \ldots, N$.

In applications, the Rayleigh criterion sets a bound on the resolution power of classical direction-finding methods. In the next sections we summarize some of the so-called super-resolution direction-finding methods which may overcome the Rayleigh bound. We divide these methods into two classes, those that use second-order and those that use second- and higher-order statistics.

3.3 Second-Order Statistics-Based Methods

The second-order methods use the sample estimate of the array spatial covariance matrix $\mathbf{R} = E\{\mathbf{r}(t)\mathbf{r}(t)^{\mathrm{H}}\} = \mathbf{A}\mathbf{R}_{\mathrm{s}}\mathbf{A}^{\mathrm{H}} + \mathbf{R}_{\mathrm{n}}$, where $\mathbf{R}_{\mathrm{s}} = E\{\mathbf{s}(t)\mathbf{s}(t)^{\mathrm{H}}\}$ is the $P \times P$ signal covariance matrix and $\mathbf{R}_{\mathrm{n}} = E\{\mathbf{n}(t)\mathbf{n}(t)^{\mathrm{H}}\}$ is the $M \times M$ noise covariance matrix. For the time being, let us assume that the noise is spatially white, i.e., $\mathbf{R}_{\mathrm{n}} = \sigma^2\mathbf{I}$. If the noise is colored and its covariance matrix is known or can be estimated, the measurements can be "whitened" by multiplying the measurements from the left by the matrix $\Lambda^{-1/2}\mathbf{E}_{\mathrm{n}}^{\mathrm{H}}$ obtained by the orthogonal eigendecomposition $\mathbf{R}_{\mathrm{n}} = \mathbf{E}_{\mathrm{n}}\Lambda\mathbf{E}_{\mathrm{n}}^{\mathrm{H}}$. The array spatial covariance matrix is estimated as $\hat{\mathbf{R}} = \sum_{t=1}^{N} \mathbf{r}(t)\mathbf{r}(t)^{\mathrm{H}}/N$.

Some spectral estimation approaches to the direction finding problem are based on optimization. Consider the "minimum variance" (MV) algorithm, for example. The received signal is processed by a beamforming vector $\mathbf{w}_{\mathrm{o}}$ which is designed such that the output power is minimized subject to the constraint that a signal from a desired direction is passed to the output with unit gain. Solving this optimization problem, we obtain the array output power as a function of the arrival angle θ as

$$P_{\mathrm{mv}}(\theta) = \frac{1}{\mathbf{a}^{\mathrm{H}}(\theta)\mathbf{R}^{-1}\mathbf{a}(\theta)}.$$

The arrival angles are obtained by scanning the range $[-90^\circ, 90^\circ]$ of θ and locating the peaks of $P_{\mathrm{mv}}(\theta)$. At low SNRs the conventional methods, such as MV, fail to resolve closely spaced arrival angles. The resolution of conventional methods are limited by SNR even if exact $\mathbf{R}$ is used, whereas in subspace

methods, there is no resolution limit; hence, the latter are also referred to as "super-resolution" methods. The limit comes from the sample estimate of $\mathbf{R}$.

The subspace-based methods exploit the eigendecomposition of the estimated array covariance matrix $\hat{\mathbf{R}}$. To see the implications of the eigendecomposition of $\hat{\mathbf{R}}$, let us first state the properties of $\mathbf{R}$: (1) If the source signals are independent or partially correlated, $\text{rank}(\mathbf{R}_s) = P$. If there are coherent sources, $\text{rank}(\mathbf{R}_s) < P$. In the methods explained in Sections 3.3.1 and 3.3.2, except for the weighted subspace fitting (WSF) method (see Section 3.3.1.1), it will be assumed that there are no coherent sources. The coherent signals case is described in Section 3.3.2. (2) If the columns of $\mathbf{A}$ are independent, which is generally true when the source bearings are different, then $\mathbf{A}$ is of full-rank P. (3) Properties 1 and 2 imply $\text{rank}(\mathbf{A}\mathbf{R}_s\mathbf{A}^H) = P$; therefore, $\mathbf{A}\mathbf{R}_s\mathbf{A}^H$ must have P nonzero eigenvalues and $M - P$ zero eigenvalues. Let the eigendecomposition of $\mathbf{A}\mathbf{R}_s\mathbf{A}^H$ be $\mathbf{A}\mathbf{R}_s\mathbf{A}^H = \sum_{i=1}^{M} \alpha_i \mathbf{e}_i \mathbf{e}_i^H$; then $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_P \geq \alpha_{P+1} = \cdots = \alpha_M = 0$ are the rank-ordered eigenvalues, and $\{\mathbf{e}_i\}_{i=1}^{M}$ are the corresponding eigenvectors. (4) Because $\mathbf{R}_n = \sigma^2\mathbf{I}$, the eigenvectors of $\mathbf{R}$ are the same as those of $\mathbf{A}\mathbf{R}_s\mathbf{A}^H$, and its eigenvalues are $\lambda_i = \alpha_i + \sigma^2$, if $1 \leq i \leq P$, or $\lambda_i = \sigma^2$, if $P + 1 \leq i \leq M$. The eigenvectors can be partitioned into two sets: $\mathbf{E}_s \triangleq [\mathbf{e}_1, \ldots, \mathbf{e}_P]$ forms the "signal subspace," whereas $\mathbf{E}_n \triangleq [\mathbf{e}_{P+1}, \ldots, \mathbf{e}_M]$ forms the "noise subspace." These subspaces are orthogonal. The signal eigenvalues $\Lambda_s \triangleq \text{diag}\{\lambda_1, \ldots, \lambda_P\}$, and the noise eigenvalues $\Lambda_n \triangleq \text{diag}\{\lambda_{P+1}, \ldots, \lambda_M\}$. (5) The eigenvectors corresponding to zero eigenvalues satisfy $\mathbf{A}\mathbf{R}_s\mathbf{A}^H\mathbf{e}_i = 0$, $i = P + 1, \ldots, M$; hence, $\mathbf{A}^H\mathbf{e}_i = 0$, $i = P + 1, \ldots, M$, because $\mathbf{A}$ and $\mathbf{R}_s$ are full rank. This last equation means that steering vectors are orthogonal to noise subspace eigenvectors. It further implies that because of the orthogonality of signal and noise subspaces, spans of signal eigenvectors and steering vectors are equal. Consequently there exists a nonsingular $P \times P$ matrix $\mathbf{T}$ such that $\mathbf{E}_s = \mathbf{A}\mathbf{T}$.

Alternatively, the signal and noise subspaces can also be obtained by performing a singular value decomposition (SVD) directly on the received data without having to calculate the array covariance matrix. Li and Vaccaro [17] state that the properties of the bearing estimates do not depend on which method is used; however, SVD must then deal with a data matrix that increases in size as the new snapshots are received. In the sequel, we assume that the array covariance matrix is estimated from the data and an eigendecomposition is performed on the estimated covariance matrix.

The eigenvalue decomposition of the spatial array covariance matrix, and the eigenvector partitionment into signal and noise subspaces, leads to a number of subspace-based direction-finding methods. The signal subspace contains information about where the signals are whereas the noise subspace informs us where they are not. Use of either subspace results in better resolution performance than conventional methods. In practice, the performance of the subspace-based methods is limited fundamentally by the accuracy of separating the two subspaces when the measurements are noisy [18]. These methods can be broadly classified into signal subspace and noise subspace methods. A summary of direction-finding methods based on both approaches is discussed in the following.

3.3.1 Signal Subspace Methods

In these methods, only the signal subspace information is retained. Their rationale is that by discarding the noise subspace we effectively enhance the SNR because the contribution of the noise power to the covariance matrix is eliminated. Signal subspace methods are divided into search-based and algebraic methods, which are explained in Sections 3.3.1.1 and 3.3.1.2.

3.3.1.1 Search-Based Methods

In search-based methods, it is assumed that the response of the array to a single source, "the array manifold" $\mathbf{a}(\theta)$, is either known analytically as a function of arrival angle, or is obtained through the calibration of the array. For example, for an M-element uniform linear array, the array response to a signal from angle θ is analytically known and is given by

$$\mathbf{a}(\theta) = \left[1, e^{-j2\pi\frac{d}{\lambda}\sin(\theta)}, \ldots, e^{-j2\pi(M-1)\frac{d}{\lambda}\sin(\theta)}\right]^T,$$

where

d is the separation between the elements

λ is the wavelength

In search-based methods to follow (except for the subspace fitting [SSF] methods), which are spatial versions of widely known power spectral density estimators, the estimated array covariance matrix is approximated by its signal subspace eigenvectors, or its "principal components," as $\hat{\mathbf{R}} \approx \sum_{i=1}^{P} \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{H}}$. Then the arrival angles are estimated by locating the peaks of a function, $S(\theta)$ $(-90^\circ \leq \theta \leq 90^\circ)$, which depends on the particular method. Some of these methods and the associated function $S(\theta)$ are summarized in the following [13,18,20]:

Correlogram method: In this method, $S(\theta) = \mathbf{a}(\theta)^{\mathrm{H}} \hat{\mathbf{R}} \mathbf{a}(\theta)$. The resolution obtained from the Correlogram method is lower than that obtained from the MV and autoregressive (AR) methods.

Minimum variance [1] *method*: In this method, $S(\theta) = 1/\mathbf{a}(\theta)^{\mathrm{H}} \hat{\mathbf{R}}^{-1} \mathbf{a}(\theta)$. The MV method is known to have a higher resolution than the correlogram method, but lower resolution and variance than the AR method.

Autoregressive method: In this method, $S(\theta) = 1/|\mathbf{u}^{\mathrm{T}} \hat{\mathbf{R}}^{-1} \mathbf{a}(\theta)|^2$ where $\mathbf{u} = [1, 0, \ldots, 0]^{\mathrm{T}}$. This method is known to have a better resolution than the previous ones.

Subspace fitting and weighted subspace fitting methods: In Section 3.2 we saw that the spans of signal eigenvectors and steering vectors are equal; therefore, bearings can be solved from the best least-squares (LS) fit of the two spanning sets when the array is calibrated [35]. In the SSF method the criterion $[\hat{\theta}, \hat{\mathbf{T}}] = \mathrm{argmin} \left\| \mathbf{E}_{\mathrm{s}} \mathbf{W}^{1/2} - \mathbf{A}(\theta)\mathbf{T} \right\|^2$ is used, where $\|.\|$ denotes the Frobenius norm, $\mathbf{W}$ is a positive definite weighting matrix, $\mathbf{E}_{\mathrm{s}}$ is the matrix of signal subspace eigenvectors, and the notation for the steering matrix is changed to show its dependence on the bearing vector θ. This criterion can be minimized directly with respect to $\mathbf{T}$, and the result for $\mathbf{T}$ can then be substituted back into it, so that

$$\hat{\theta} = \mathrm{argmin}\ \mathrm{Tr}\left\{ (\mathbf{I} - \mathbf{A}(\theta)\mathbf{A}(\theta)^{\#}) \mathbf{E}_{\mathrm{s}} \mathbf{W} \mathbf{E}_{\mathrm{s}}^{\mathrm{H}} \right\},$$

where $\mathbf{A}^{\#} = (\mathbf{A}^{\mathrm{H}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{H}}$.

Viberg and Ottersten have shown that a class of direction finding algorithms can be approximated by this SSF formulation for appropriate choices of the weighting matrix $\mathbf{W}$. For example, for the deterministic ML method $\mathbf{W} = \Lambda_{\mathrm{s}} - \sigma^2\mathbf{I}$, which is implemented using the empirical values of the signal eigenvalues, Λ_{s}, and the noise eigenvalue σ^2. Total least square (TLS)-estimation of signal parameters via rotational invariance techniques (ESPRIT), which is explained in the next section, can also be formulated in a similar but more involved way. Viberg and Ottersten have also derived an optimal WSF method, which yields the smallest estimation error variance among the class of SSF methods. In WSF, $\mathbf{W} = (\Lambda_{\mathrm{s}} - \sigma^2\mathbf{I})^2 \Lambda_{\mathrm{s}}^{-1}$. The WSF method works regardless of the source covariance (including coherence) and has been shown to have the same asymptotic properties as the stochastic ML method; hence, it is asymptotically efficient for Gaussian signals (i.e., it achieves the stochastic CRB). Its behavior in the finite sample case may be different from the asymptotic case [34]. Viberg and Ottersten have also shown that the asymptotic properties of the WSF estimates are identical for both cases of Gaussian and non-Gaussian sources. They have also developed a consistent detection method for arbitrary signal correlation, and an algorithm for minimizing the WSF criterion. They do point out several practical implementation problems of their method, such as the need for accurate calibrations of the array manifold and knowledge of the derivative of the steering vectors w.r.t. θ. For nonlinear and nonuniform arrays, MD search methods are required for SSF, hence it is computationally expensive.

3.3.1.2 Algebraic Methods

Algebraic methods do not require a search procedure and yield DOA estimates directly.

ESPRIT [23]: The ESPRIT algorithm requires "translationally invariant" arrays, i.e., an array with its "identical copy" displaced in space. The geometry and response of the arrays do not have to be known; only the measurements from these arrays and the displacement between the identical arrays are required. The computational complexity of ESPRIT is less than that of the search-based methods.

Let $\mathbf{r}^1(t)$ and $\mathbf{r}^2(t)$ be the measurements from these arrays. Due to the displacement of the arrays the following holds

$$\mathbf{r}^1(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}_1(t) \quad \text{and} \quad \mathbf{r}^2(t) = \mathbf{A}\Phi\mathbf{s}(t) + \mathbf{n}_2(t),$$

where $\Phi = \text{diag}\left\{e^{-j2\pi\frac{d}{\lambda}\sin\theta_1}, \ldots, e^{-j2\pi\frac{d}{\lambda}\sin\theta_P}\right\}$ in which d is the separation between the identical arrays, and the angles $\{\theta_i\}_{i=1}^P$ are measured with respect to the normal to the displacement vector between the identical arrays. Note that the auto covariance of $\mathbf{r}^1(t)$, $\mathbf{R}^{11}$, and the cross-covariance between $\mathbf{r}^1(t)$ and $\mathbf{r}^2(t)$, $\mathbf{R}^{21}$, are given by

$$\mathbf{R}^{11} = \mathbf{A}\mathbf{D}\mathbf{A}^{\text{H}} + \mathbf{R}_{\text{n}_1},$$

and

$$\mathbf{R}^{21} = \mathbf{A}\Phi\mathbf{D}\mathbf{A}^{\text{H}} + \mathbf{R}_{\text{n}_2\text{n}_1},$$

where

$\mathbf{D}$ is the covariance matrix of the sources

$\mathbf{R}_{\text{n}_1}$ and $\mathbf{R}_{\text{n}_2\text{n}_1}$ are the noise auto- and cross-covariance matrices

The ESPRIT algorithm solves for Φ, which then gives the bearing estimates. Although the subspace separation concept is not used in ESPRIT, its LS and TLS versions are based on a signal subspace formulation. The LS and TLS versions are more complicated, but are more accurate than the original ESPRIT, and are summarized in the next subsection. Here we summarize the original ESPRIT:

1. Estimate the autocovariance of $\mathbf{r}^1(t)$ and cross covariance between $\mathbf{r}^1(t)$ and $\mathbf{r}^2(t)$, as

$$\mathbf{R}^{11} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{r}^1(t)\mathbf{r}^1(t)^{\text{H}},$$

   and

$$\mathbf{R}^{21} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{r}^2(t)\mathbf{r}^1(t)^{\text{H}}.$$

2. Calculate $\hat{\mathbf{R}}^{11} = \mathbf{R}^{11} - \mathbf{R}_{\text{n}_1}$ and $\hat{\mathbf{R}}^{21} = \mathbf{R}^{21} - \mathbf{R}_{\text{n}_2\text{n}_1}$, where $\mathbf{R}_{\text{n}_1}$ and $\mathbf{R}_{\text{n}_2\text{n}_1}$ are the estimated noise covariance matrices.
3. Find the singular values λ_i of the matrix pencil $\hat{\mathbf{R}}^{11} - \lambda_i\hat{\mathbf{R}}^{21}$, $i = 1, \ldots, P$.
4. The bearings, θ_i $(i = 1, \ldots, P)$, are readily obtained by solving the equation

$$\lambda_i = e^{j2\pi\frac{d}{\lambda}\sin\theta_i},$$

   for θ_i. In the above steps, it is assumed that the noise is spatially and temporally white or the covariance matrices $\mathbf{R}_{\text{n}_1}$ and $\mathbf{R}_{\text{n}_2\text{n}_1}$ are known.

LS and TLS-ESPRIT [28]:

1. Follow Steps 1 and 2 of ESPRIT.
2. Stack $\hat{\mathbf{R}}^{11}$ and $\hat{\mathbf{R}}^{21}$ into a $2M \times M$ matrix $\mathbf{R}$, as $\mathbf{R} \stackrel{\Delta}{=} [\hat{\mathbf{R}}^{11T} \hat{\mathbf{R}}^{21T}]^{\mathrm{T}}$, and perform an SVD of $\mathbf{R}$, keeping the first $2M \times P$ submatrix of the left singular vectors of $\mathbf{R}$. Let this submatrix be $\mathbf{E}_s$.
3. Partition $\mathbf{E}_s$ into two $M \times P$ matrices $\mathbf{E}_{s1}$ and $\mathbf{E}_{s2}$ such that

$$\mathbf{E}_s = \left[\mathbf{E}_{s1}^{\mathrm{T}} \mathbf{E}_{s2}^{\mathrm{T}}\right]^{\mathrm{T}}.$$

4. For LS-ESPRIT, calculate the eigendecomposition of $\left(\mathbf{E}_{s1}^{\mathrm{H}} \mathbf{E}_{s1}\right)^{-1} \mathbf{E}_{s1}^{\mathrm{H}} \mathbf{E}_{s2}$. The eigenvalue matrix gives

$$\Phi = \mathrm{diag}\left\{ \mathrm{e}^{-j2\pi \frac{d}{\lambda} \sin\theta_1}, \ldots, \mathrm{e}^{-j2\pi \frac{d}{\lambda} \sin\theta_P} \right\},$$

   from which the arrival angles are readily obtained. For TLS-ESPRIT, proceed as follows.
5. Perform an SVD of the $M \times 2P$ matrix $[\mathbf{E}_{s1}, \mathbf{E}_{s2}]$, and stack the last P right singular vectors of $[\mathbf{E}_{s1}, \mathbf{E}_{s2}]$ into a $2P \times P$ matrix denoted $\mathbf{F}$.
6. Partition $\mathbf{F}$ as

$$\mathbf{F} \stackrel{\Delta}{=} \left[\mathbf{F}_x^{\mathrm{T}} \mathbf{F}_y^{\mathrm{T}}\right]^{\mathrm{T}},$$

   where $\mathbf{F}_x$ and $\mathbf{F}_y$ are $P \times P$.
7. Perform the eigendecomposition of $-\mathbf{F}_x \mathbf{F}_y^{-1}$. The eigenvalue matrix gives

$$\Phi = \mathrm{diag}\left\{ \mathrm{e}^{-j2\pi \frac{d}{\lambda} \sin\theta_1}, \ldots, \mathrm{e}^{-j2\pi \frac{d}{\lambda} \sin\theta_P} \right\},$$

   from which the arrival angles are readily obtained.

Different versions of ESPRIT have different statistical properties. The Toeplitz approximation method (TAM) [16], in which the array measurement model is represented as a state-variable model, although different in implementation from LS-ESPRIT, is equivalent to LS-ESPRIT; hence, it has the same error variance as LS-ESPRIT.

Generalized eigenvalues utilizing signal subspace eigenvectors (GEESE) [24]

1. Follow Steps 1 through 3 of TLS-ESPRIT.
2. Find the singular values λ_i of the pencil

$$\mathbf{E}_{s1} - \lambda_i \mathbf{E}_{s2}, \quad i = 1, \ldots, P.$$

3. The bearings, θ_i $(i = 1, \ldots, P)$, are readily obtained from

$$\lambda_i = \mathrm{e}^{j2\pi \frac{d}{\lambda} \sin\theta_i}.$$

   The GEESE method is claimed to be better than ESPRIT [24].

3.3.2 Noise Subspace Methods

These methods, in which only the noise subspace information is retained, are based on the property that the steering vectors are orthogonal to any linear combination of the noise subspace eigenvectors. Noise subspace methods are also divided into search-based and algebraic methods, which are explained next.

3.3.2.1 Search-Based Methods

In search-based methods, the array manifold is assumed to be known, and the arrival angles are estimated by locating the peaks of the function $S(\theta) = 1/\mathbf{a}(\theta)^{\mathrm{H}}\mathbf{N}\mathbf{a}(\theta)$, where $\mathbf{N}$ is a matrix formed using the noise space eigenvectors.

Pisarenko method: In this method, $\mathbf{N} = \mathbf{e}_M\mathbf{e}_M^{\mathrm{H}}$, where $\mathbf{e}_M$ is the eigenvector corresponding to the minimum eigenvalue of $\mathbf{R}$. If the minimum eigenvalue is repeated, any unit-norm vector which is a linear combination of the eigenvectors corresponding to the minimum eigenvalue can be used as $\mathbf{e}_M$. The basis of this method is that when the search angle θ corresponds to an actual arrival angle, the denominator of $S(\theta)$ in the Pisarenko method, $\left|\mathbf{a}(\theta)^{\mathrm{H}}\mathbf{e}_M\right|^2$, becomes small due to orthogonality of steering vectors and noise subspace eigenvectors; hence, $S(\theta)$ will peak at an arrival angle.

Multiple signal classification (MUSIC) [29] *method*: In this method, $\mathbf{N} = \sum_{i=P+1}^{M}\mathbf{e}_i\mathbf{e}_i^{\mathrm{H}}$. The idea is similar to that of the Pisarenko method; the inner product $\left|\mathbf{a}(\theta)^{\mathrm{H}}\sum_{i=P+1}^{M}\mathbf{e}_i\right|^2$ is small when θ is an actual arrival angle. An obvious signal-subspace formulation of MUSIC is also possible. The MUSIC spectrum is equivalent to the MV method using the exact covariance matrix when SNR is infinite, and therefore performs better than the MV method.

Asymptotic properties of MUSIC are well established [32,33], e.g., MUSIC is known to have the same asymptotic variance as the deterministic ML method for uncorrelated sources. It is shown by Xu and Buckley [38] that although, asymptotically, bias is insignificant compared to standard deviation, it is an important factor limiting the performance for resolving closely spaced sources when they are correlated.

In order to overcome the problems due to finite sample effects and source correlation, a MD version of MUSIC has been proposed [28,29]; however, this approach involves a computationally involved search, as in the ML method. MD MUSIC can be interpreted as a norm minimization problem, as shown in Ephraim et al. [8]; using this interpretation, strong consistency of MD MUSIC has been demonstrated. An optimally weighted version of MD MUSIC, which outperforms the deterministic ML method, has also been proposed in Viberg and Ottersten [35].

Eigenvector (EV) method: In this method,

$$\mathbf{N} = \sum_{i=P+1}^{M}\frac{1}{\lambda_i}\mathbf{e}_i\mathbf{e}_i^{\mathrm{H}}.$$

The only difference between the EV method and MUSIC is the use of inverse eigenvalue (the λ_i are the noise subspace eigenvalues of $\mathbf{R}$) weighting in eigenvector and unity weighting in MUSIC, which causes eigenvector to yield fewer spurious peaks than MUSIC [13]. The EV method is also claimed to shape the noise spectrum better than MUSIC.

Method of direction estimation (MODE): MODE is equivalent to WSF when there are no coherent sources. Viberg and Ottersten [35] claim that, for coherent sources, only WSF is asymptotically efficient. A minimum-norm interpretation and proof of strong consistency of MODE for ergodic and stationary signals, has also been reported [8]. The norm measure used in that work involves the source covariance matrix. By contrasting this norm with the Frobenius norm that is used in MD MUSIC, Ephraim et al. relate MODE and MD MUSIC.

Minimum-norm [15] *method*: In this method, the matrix $\mathbf{N}$ is obtained as follows [12]:

1. Form $\mathbf{E}_n = [\mathbf{e}_{P+1}, \ldots, \mathbf{e}_M]$.
2. Partition $\mathbf{E}_n$ as $\mathbf{E}_n = [\mathbf{c}\mathbf{C}^{\mathrm{T}}]^{\mathrm{T}}$, to establish $\mathbf{c}$ and $\mathbf{C}$.
3. Compute $\mathbf{d} = [1(\,(\mathbf{c}^{\mathrm{H}}\mathbf{c})^{-1}\mathbf{C}^{*}\mathbf{c})^{\mathrm{T}}]^{\mathrm{T}}$, and, finally, $\mathbf{N} = \mathbf{d}\mathbf{d}^{\mathrm{H}}$.

For two closely spaced, equal power signals, the minimum-norm method has been shown to have a lower SNR threshold (i.e., the minimum SNR required to separate the two sources) than MUSIC [14].

Li and Vaccaro [17] derive and compare the mean-squared errors of the DOA estimates from minimum-norm and MUSIC algorithms due to finite sample effects, calibration errors, and noise modeling errors for the case of finite samples and high SNR. They show that mean-squared errors for DOA estimates produced by the MUSIC algorithm are always lower than the corresponding mean-squared errors for the minimum-norm algorithm.

3.3.2.2 Algebraic Methods

When the array is uniform linear, so that

$$\mathbf{a}(\theta) = \left[1, \mathrm{e}^{-j2\pi\frac{d}{\lambda}\sin(\theta)}, \ldots, \mathrm{e}^{-j2\pi(M-1)\frac{d}{\lambda}\sin(\theta)}\right]^{\mathrm{T}},$$

the search in $S(\theta) = 1/\mathbf{a}(\theta)^{\mathrm{H}}\mathbf{N}\mathbf{a}(\theta)$ for the peaks can be replaced by a root-finding procedure which yields the arrival angles. So doing results in better resolution than the search-based alternative because the root-finding procedure can give distinct roots corresponding to each source whereas the search function may not have distinct maxima for closely spaced sources. In addition, the computational complexity of algebraic methods is lower than that of the search-based ones. The algebraic version of MUSIC (root-MUSIC) is given next; for algebraic versions of Pisarenko, EV, and minimum-norm, the matrix $\mathbf{N}$ in root-MUSIC is replaced by the corresponding $\mathbf{N}$ in each of these methods.

Root-MUSIC method: In root-MUSIC, the array is required to be uniform linear, and the search procedure in MUSIC is converted into the following root-finding approach:

1. Form the $M \times M$ matrix $\mathbf{N} = \sum_{i=P+1}^{M} \mathbf{e}_i\mathbf{e}_i^{\mathrm{H}}$.
2. Form a polynomial $p(z)$ of degree $2M-1$ which has for its ith coefficient $c_i = \mathrm{tr}_i[\mathbf{N}]$, where tr_i denotes the trace of the ith diagonal, and $i = -(M-1), \ldots, 0, \ldots, M-1$. Note that tr_0 denotes the main diagonal, tr_1 denotes the first super-diagonal, and tr_{-1} denotes the first sub-diagonal.
3. The roots of $p(z)$ exhibit inverse symmetry with respect to the unit circle in the z-plane. Express $p(z)$ as the product of two polynomials $p(z) = h(z)h^{\star}(z^{-1})$.
4. Find the roots z_i $(i = 1, \ldots, M)$ of $h(z)$. The angles of roots that are very close to (or, ideally on) the unit circle yield the DOA estimates, as

$$\theta_i = \sin^{-1}\left(\frac{\lambda}{2\pi d}\angle z_i\right), \quad \text{where } i = 1, \ldots, P.$$

The root-MUSIC algorithm has been shown to have better resolution power than MUSIC [27]; however, as mentioned previously, root-MUSIC is restricted to uniform linear arrays (ULA). Steps 2 through 4 make use of this knowledge. Li and Vaccaro show that algebraic versions of the MUSIC and minimum-norm algorithms have the same mean-squared errors as their search-based versions for finite samples and high SNR case. The advantages of root-MUSIC over search-based MUSIC is increased resolution of closely spaced sources and reduced computations.

3.3.3 Spatial Smoothing

When there are coherent (completely correlated) sources, rank($\mathbf{R}_{\mathrm{s}}$), and consequently rank($\mathbf{R}$), is less than P, and hence the above described subspace methods fail. If the array is uniform linear, then by applying the spatial smoothing method, described below, a new rank-P matrix is obtained which can be used in place of $\mathbf{R}$ in any of the subspace methods described earlier.

Spatial smoothing [9,31] starts by dividing the M-vector $\mathbf{r}(t)$ of the ULA into $K = M - S + 1$ overlapping subvectors of size S, $\mathbf{r}^{\mathrm{f}}_{S,k}(k = 1, \ldots, K)$, with elements $\{r_k, \ldots, r_{k+S-1}\}$, and $\mathbf{r}^{\mathrm{b}}_{S,k}(k = 1, \ldots, K)$, with elements $\{r^{*}_{M-k+1}, \ldots, r^{*}_{M-S-k+2}\}$. Then, a forward and backward spatially smoothed matrix $\mathbf{R}^{\mathrm{fb}}$ is calculated as

$$\mathbf{R}^{\mathrm{fb}} = \sum_{t=1}^{N} \sum_{k=1}^{K} \left(\mathbf{r}^{\mathrm{f}}_{S,k}(t) \mathbf{r}^{\mathrm{f}\,\mathrm{H}}_{S,k}(t) + \mathbf{r}^{\mathrm{b}}_{S,k}(t) \mathbf{r}^{\mathrm{b}\,\mathrm{H}}_{S,k}(t) \right) / KN.$$

The rank of $\mathbf{R}^{\mathrm{fb}}$ is P if there are at most $2M/3$ coherent sources. S must be selected such that

$$P_{\mathrm{c}} + 1 \leq S \leq M - P_{\mathrm{c}}/2 + 1,$$

in which P_{c} is the number of coherent sources. Then, any subspace-based method can be applied to $\mathbf{R}^{\mathrm{fb}}$ to determine the DOA. It is also possible to do spatial smoothing based only on $\mathbf{r}^{\mathrm{f}}_{S,k}$ or $\mathbf{r}^{\mathrm{b}}_{S,k}$, but in this case at most $M/2$ coherent sources can be handled.

3.3.4 Discussion

The application of all the subspace-based methods requires exact knowledge of the number of signals, in order to separate the signal and noise subspaces. The number of signals can be estimated from the data using either the Akaike information criterion (AIC) [36] or minimum descriptive length (MDL) [37] methods. The effect of underestimating the number of sources is analyzed by Radich and Buckley [26], whereas the case of overestimating the number of signals can be treated as a special case of the analysis in Stoica and Nehorai [32].

The second-order methods described above have the following disadvantages:

1. Except for ESPRIT (which requires a special array structure), all of the above methods require calibration of the array which means that the response of the array for every possible combination of the source parameters should be measured and stored; or, analytical knowledge of the array response is required. However, at any time, the antenna response can be different from when it was last calibrated due to environmental effects such as weather conditions for radar, or water waves for sonar. Even if the analytical response of the array elements is known, it may be impossible to know or track the precise locations of the elements in some applications (e.g., towed array). Consequently, these methods are sensitive to errors and perturbations in the array response. In addition, physically identical sensors may not respond identically in practice due to lack of synchronization or imbalances in the associated electronic circuitry.
2. In deriving the above methods, it was assumed that the noise covariance structure is known; however, it is often unrealistic to assume that the noise statistics are known due to several reasons. In practice, the noise is not isolated; it is often observed along with the signals. Moreover, as Swindlehurst and Kailath [33] state, there are noise phenomena effects that cannot be modeled accurately, e.g., channel crosstalk, reverberation, near-field, wideband, and distributed sources.
3. None of the methods in Sections 3.3.1 and 3.3.2, except for the WSF method and other MD search-based approaches, which are computationally very expensive, work when there are coherent (completely correlated) sources. Only if the array is uniform linear, can the spatial smoothing method in Section 3.3.2 be used. On the other hand, higher-order statistics of the received signals can be exploited to develop direction-finding methods which have less restrictive requirements.

3.4 Higher-Order Statistics-Based Methods

The higher-order statistical direction-finding methods use the spatial cumulant matrices of the array. They require that the source signals be non-Gaussian so that their higher than second-order statistics convey extra information. Most communication signals (e.g., Quadrature Amplitude Modulation (QAM)) are "complex circular" (a signal is complex circular if its real and imaginary parts are independent and symmetrically distributed with equal variances) and hence their third-order cumulants vanish; therefore, even-order cumulants are used, and usually fourth-order cumulants are employed. The fourth-order cumulant of the source signals must be nonzero in order to use these methods. One important feature of cumulant-based methods is that they can suppress Gaussian noise regardless of its coloring. Consequently, the requirement of having to estimate the noise covariance, as in second-order statistical processing methods, is avoided in cumulant-based methods. It is also possible to suppress non-Gaussian noise [6], and, when properly applied, cumulants extend the aperture of an array [5,30], which means that more sources than sensors can be detected. As in the second-order statistics-based methods, it is assumed that the number of sources is known or is estimated from the data.

The fourth-order moments of the signal $\mathbf{s}(t)$ are

$$E\{s_i s_j^* s_k s_l^*\}, \quad 1 \le i, j, k, l \le P,$$

and the fourth-order cumulants are defined as

$$\begin{aligned} c_{4,s}(i,j,k,l) &= \text{cum}(s_i, s_j^*, s_k, s_l^*) \\ &= E\{s_i s_j^* s_k s_l^*\} - E\{s_i s_j^*\}E\{s_k s_l^*\} - E\{s_i s_l^*\}E\{s_k s_j^*\} - E\{s_i s_j\}E\{s_k^* s_l^*\}, \end{aligned}$$

where $1 \le i, j, k, l \le P$. Note that two arguments in the above fourth-order moments and cumulants are conjugated and the other two are unconjugated. For circularly symmetric signals, which is often the case in communication applications, the last term in $c_{4,s}(i,j,k,l)$ is zero.

In practice, sample estimates of the cumulants are used in place of the theoretical cumulants, and these sample estimates are obtained from the received signal vector $\mathbf{r}(t)$ $(t = 1, \ldots, N)$, as

$$\begin{aligned} \hat{c}_{4,r}(i,j,k,l) = &\sum_{t=1}^{N} r_i(t) r_j^*(t) r_k(t) r_l^*(t)/N - \sum_{t=1}^{N} r_i(t) r_j^*(t) \sum_{t=1}^{N} r_k(t) r_l^*(t)/N^2 \\ &- \sum_{t=1}^{N} r_i(t) r_l^*(t) \sum_{t=1}^{N} r_k(t) r_j^*(t)/N^2, \end{aligned}$$

where $1 \le i, j, k, l \le M$. Note that the last term in $c_{4,r}(i,j,k,l)$ is zero and, therefore, it is omitted.

Higher-order statistical subspace methods use fourth-order spatial cumulant matrices of the array output, which can be obtained in a number of ways by suitably selecting the arguments i, j, k, l of $c_{4,r}(i,j,k,l)$. Existing methods for the selection of the cumulant matrix, and their associated processing schemes are summarized next.

Pan–Nikias [22] *and Cardoso–Moulines* [2] *method*: In this method, the array needs to be calibrated, or its response must be known in analytical form. The source signals are assumed to be independent or partially correlated (i.e., there are no coherent signals). The method is as follows:

1. An estimate of an $M \times M$ fourth-order cumulant matrix $\mathbf{C}$ is obtained from the data. The following two selections for $\mathbf{C}$ are possible [2,22]:

$$c_{ij} = c_{4,r}(i,j,j,j), \quad 1 \le i, j \le M,$$

or

$$c_{ij} = \sum_{m=1}^{M} c_{4,r}(i,j,m,m), \quad 1 \le i,j \le M.$$

Using cumulant properties [19], and Equation 3.1, and a_{ij} for the ijth element of $\mathbf{A}$, it is easy to verify that

$$c_{4,r}(i,j,j,j) = \sum_{p=1}^{P} a_{ip} \sum_{q,r,s=1}^{P} a_{jq}^{\star} a_{jr} a_{js}^{\star} c_{4,s}(p,q,r,s),$$

which, in matrix format, is $\mathbf{C} = \mathbf{AB}$ where $\mathbf{A}$ is the steering matrix and $\mathbf{B}$ is a $P \times M$ matrix with elements

$$b_{ij} = \sum_{q,r,s=1}^{P} a_{iq}^{\star} a_{jr} a_{js}^{\star} c_{4,s}(i,q,r,s).$$

Similarly,

$$\sum_{m=1}^{M} c_{4,r}(i,j,m,m) = \sum_{p,q=1}^{P} a_{ip} \left(\sum_{r,s=1}^{P} \sum_{m=1}^{M} a_{mr} a_{ms}^{\star} c_{4,s}(p,q,r,s) \right) a_{jq}^{\star}, \quad 1 \le i,j \le M,$$

which, in matrix form, can be expressed as $\mathbf{C} = \mathbf{ADA}^{\mathrm{H}}$, where $\mathbf{D}$ is a $P \times P$ matrix with elements

$$d_{ij} = \sum_{r,s=1}^{P} \sum_{m=1}^{M} a_{mr} a_{ms}^{\star} c_{4,,s}(i,j,r,s).$$

Note that additive Gaussian noise is suppressed in both C matrices because higher than second-order statistics of a Gaussian process are zero.

2. The P left singular vectors of $\mathbf{C} = \mathbf{AB}$, corresponding to nonzero singular values or the P eigenvectors of $\mathbf{C} = \mathbf{ADA}^{\mathrm{H}}$ corresponding to nonzero eigenvalues form the signal subspace. The orthogonal complement of the signal subspace gives the noise subspace. Any of the Section 3.3 covariance-based search and algebraic direction finding (DF) methods (except for the EV method and ESPRIT) can now be applied (in exactly the same way as described in Section 3.3) either by replacing the signal and noise subspace eigenvectors and eigenvalues of the array covariance matrix by the corresponding subspace eigenvectors and eigenvalues of $\mathbf{ADA}^{\mathrm{H}}$, or by the corresponding subspace singular vectors and singular values of $\mathbf{AB}$. A cumulant-based analog of the EV method does not exist because the eigenvalues and singular values of $\mathbf{ADA}^{\mathrm{H}}$ and $\mathbf{AB}$ corresponding to the noise subspace are theoretically zero. The cumulant-based analog of ESPRIT is explained later.

The same assumptions and restrictions for the covariance-based methods apply to their analogs in the cumulant domain. The advantage of using the cumulant-based analogs of these methods is that there is no need to know or estimate the noise-covariance matrix.

The asymptotic covariance of the DOA estimates obtained by MUSIC based on the above fourth-order cumulant matrices are derived in Cardoso and Moulines [2] for the case of Gaussian measurement noise with arbitrary spatial covariance, and are compared to the asymptotic covariance of the DOA estimates

from the covariance-based MUSIC algorithm. Cardoso and Moulines show that covariance- and fourth-order cumulant-based MUSIC have similar performance for the high SNR case, and as SNR decreases below a certain SNR threshold, the variances of the fourth-order cumulant-based MUSIC DOA estimates increase with the fourth power of the reciprocal of the SNR, whereas the variances of covariance-based MUSIC DOA estimates increase with the square of the reciprocal of the SNR. They also observe that for high SNR and uncorrelated sources, the covariance-based MUSIC DOA estimates are uncorrelated, and the asymptotic variance of any particular source depends only on the power of that source (i.e., it is independent of the powers of the other sources). They observe, on the other hand, that DOA estimates from cumulant-based MUSIC, for the same case, are correlated, and the variance of the DOA estimate of a weak source increases in the presence of strong sources. This observation limits the use of cumulant-based MUSIC when the sources have a high dynamic range, even for the case of high SNR. Cardoso and Moulines state that this problem may be alleviated when the source of interest has a large fourth-order cumulant.

Porat and Friedlander [25] *method*: In this method, the array also needs to be calibrated, or its response is required in analytical form. The model used in this method divides the sources into groups that are partially correlated (but not coherent) within each group, but are statistically independent across the groups, i.e.,

$$\mathbf{r}(t) = \sum_{g=1}^{G} \mathbf{A}_g \mathbf{s}_g + \mathbf{n}(t),$$

where G is the number of groups each having p_g sources $\left(\sum_{g=1}^{G} p_g = P\right)$. In this model, the p_g sources in the gth group are partially correlated, and they are received from different directions. The method is as follows:

1. Estimate the fourth-order cumulant matrix, $\mathbf{C}_r$, of $\mathbf{r}(t) \otimes \mathbf{r}(t)^*$, where $\otimes$ denotes the Kronecker product. It can be verified that

$$\mathbf{C}_r = \sum_{g=1}^{G} (\mathbf{A}_g \otimes \mathbf{A}_g^*) \mathbf{C}_{s_g} (\mathbf{A}_g \otimes \mathbf{A}_g^*)^{\mathrm{H}},$$

   where $\mathbf{C}_{s_g}$ is the fourth-order cumulant matrix of $\mathbf{s}_g$. The rank of $\mathbf{C}_r$ is $\sum_{g=1}^{G} p_g^2$, and since $\mathbf{C}_r$ is $M^2 \times M^2$, it has $M^2 - \sum_{g=1}^{G} p_g^2$ zero eigenvalues which correspond to the noise subspace. The other eigenvalues correspond to the signal subspace.
2. Compute the SVD of $\mathbf{C}_r$ and identify the signal and noise subspace singular vectors. Now, second-order subspace-based search methods can be applied, using the signal or noise subspaces, by replacing the array response vector $\mathbf{a}(\theta)$ by $\mathbf{a}(\theta) \otimes \mathbf{a}^*(\theta)$.

The eigendecomposition in this method has computational complexity $O(M^6)$ due to the Kronecker product, whereas the second-order statistics-based methods (e.g., MUSIC) have complexity $O(M^3)$.

Chiang–Nikias [4] *method*: This method uses the ESPRIT algorithm and requires an array with its entire identical copy displaced in space by distance d; however, no calibration of the array is required. The signals

$$\mathbf{r}^1(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}_1(t),$$

and

$$\mathbf{r}^2(t) = \mathbf{A}\Phi\mathbf{s}(t) + \mathbf{n}_2(t).$$

Two $M \times M$ matrices $\mathbf{C}^1$ and $\mathbf{C}^2$ are generated as follows:

$$c_{ij}^1 = \text{cum}\left(r_i^1, r_j^{1\star}, r_k^1, r_k^{1\star}\right), \quad 1 \le i, j, k \le M,$$

and

$$c_{ij}^2 = \text{cum}\left(r_i^2, r_j^{1\star}, r_k^1, r_k^{1\star}\right), \quad 1 \le i, j, k \le M.$$

It can be shown that $\mathbf{C}^1 = \mathbf{AEA}^{\text{H}}$ and $\mathbf{C}^2 = \mathbf{A\Phi EA}^{\text{H}}$, where

$$\Phi = \text{diag}\left\{e^{-j2\pi\frac{d}{\lambda}\sin\theta_1}, \ldots, e^{-j2\pi\frac{d}{\lambda}\sin\theta_P}\right\},$$

in which d is the separation between the identical arrays, and $\mathbf{E}$ is a $P \times P$ matrix with elements

$$e_{ij} = \sum_{q,r=1}^{P} a_{kq} a_{kr}^{\star} c_{4,s}(i, q, r, j).$$

Note that these equations are in the same form as those for covariance-based ESPRIT (the noise cumulants do not appear in $\mathbf{C}^1$ and $\mathbf{C}^2$ because the fourth-order cumulants of Gaussian noises are zero); therefore, any version of ESPRIT or GEESE can be used to solve for Φ by replacing $\mathbf{R}^{11}$ and $\mathbf{R}^{21}$ by $\mathbf{C}^1$ and $\mathbf{C}^2$, respectively.

Virtual cross-correlation computer (VC^3) [5]: In VC^3, the source signals are assumed to be statistically independent. The idea of VC^3 can be demonstrated as follows: Suppose we have three identical sensors as in Figure 3.1, where $r_1(t), r_2(t)$, and $r_3(t)$ are measurements, and $\vec{d}_1$, $\vec{d}_2$, and $\vec{d}_3$ ($\vec{d}_3 = \vec{d}_1 + \vec{d}_2$) are the vectors joining these sensors. Let the response of each sensor to a signal from θ be $a(\theta)$. A "virtual" sensor is one at which no measurement is actually made. Suppose that we wish to compute the correlation between the virtual sensor $v_1(t)$ and $r_2(t)$, which (using the plane wave assumption) is

$$E\{r_2^{\star}(t) v_1(t)\} = \sum_{p=1}^{P} |a(\theta_p)|^2 \sigma_p^2 e^{-j\vec{k}_p \cdot \vec{d}_3}.$$

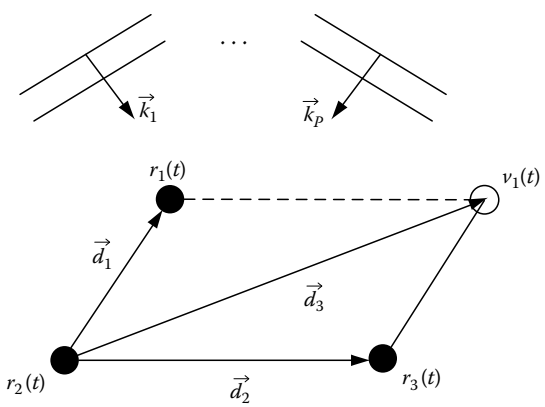

FIGURE 3.1 Demonstration of VC^3.

Consider the following cumulant

$$\begin{aligned}\text{cum}(r_2^*(t), r_1(t), r_2^*(t), r_3(t)) &= \sum_{p=1}^{P} |a(\theta_p)|^4 \gamma_p e^{-j\vec{k}_p.\vec{d}_1} e^{-j\vec{k}_p.\vec{d}_2} \\ &= \sum_{p=1}^{P} |a(\theta_p)|^4 \gamma_p e^{-j\vec{k}_p.\vec{d}_3}.\end{aligned}$$

This cumulant carries the same angular information as the cross correlation $E\{r_2^*(t)v_1(t)\}$, but for sources having different powers.

The fact that we are interested only in the directional information carried by correlations between the sensors therefore let us interpret a cross correlation as a vector (e.g., $\vec{d}_3$), and a fourth-order cumulant as the addition of two vectors (e.g., $\vec{d}_1 + \vec{d}_2$). This interpretation leads to the idea of decomposing the computation of a cross correlation into that of computing a cumulant. Doing this means that the directional information that would be obtained from the cross correlation between nonexisting sensors (or between an actual sensor and a nonexisting sensor) at certain virtual locations in the space can be obtained from a suitably defined cumulant that uses the real sensor measurements.

One advantage of virtual cross-correlation computation is that it is possible to obtain a larger aperture than would be obtained by using only second-order statistics. This means that more sources than sensors can be detected using cumulants. For example, given an M element ULA, VC^3 lets its aperture be extended from M to $2M-1$ sensors, so that $2M-2$ targets can be detected (rather than $M-1$) just by using the array covariance matrix obtained by VC^3 in any of the subspace-based search methods explained earlier. This use of VC^3 requires the array to be calibrated. Another advantage of VC^3 is a fault tolerance capability. If sensors at certain locations in a given array fail to operate properly, these sensors can be replaced using VC^3.

Virtual ESPRIT (VESPA) [5]: For VESPA, the array only needs two identical sensors; the rest of the array may have arbitrary and unknown geometry and response. The sources are assumed to be statistically independent. VESPA uses the ESPRIT solution applied to cumulant matrices. By choosing a suitable pair of cumulants in VESPA, the need for a copy of the entire array, as required in ESPRIT, is totally eliminated. VESPA preserves the computational advantage of ESPRIT over search-based algorithms. An example array configuration is given in Figure 3.2.

Without loss of generality, let the signals received by the identical sensor pair be r_1 and r_2. The sensors r_1 and r_2 are collectively referred to as the "guiding sensor pair." The VESPA algorithm is

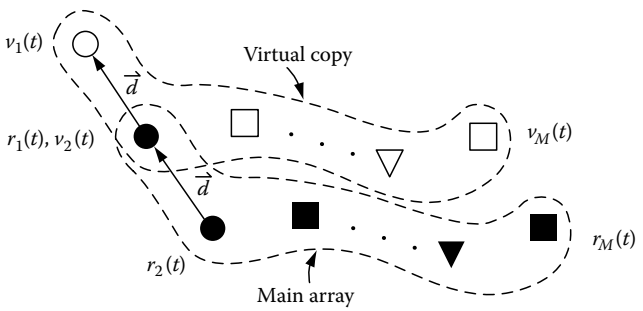

FIGURE 3.2 The main array and its virtual copy.

1. Two $M \times M$ matrices, $\mathbf{C}^1$ and $\mathbf{C}^2$, are generated as follows:

$$c_{ij}^1 = \text{cum}(r_1, r_1^*, r_i, r_j^*), \quad 1 \le i, j \le M,$$
$$c_{ij}^2 = \text{cum}(r_2, r_1^*, r_i, r_j^*), \quad 1 \le i, j \le M.$$

   It can be shown that these relations can be expressed as $\mathbf{C}^1 = \mathbf{AFA}^{\text{H}}$ and $\mathbf{C}^2 = \mathbf{A\Phi FA}^{\text{H}}$, where the $P \times P$ matrix

$$\mathbf{F} = \text{diag}\left\{\gamma_{4,s_1}|a_{11}|^2, \ldots, \gamma_{4,s_P}|a_{1P}|^2\right\}, \{\gamma_{4,s_P}\}_{p=1}^P,$$

   and Φ has been defined before.
2. Note that these equations are in the same form as ESPRIT and Chiang and Nikias's ESPRIT-like method; however, as opposed to these methods, there is no need for an identical copy of the array; only an identical response sensor pair is necessary for VESPA. Consequently, any version of ESPRIT or GEESE can be used to solve for Φ by replacing $\mathbf{R}^{11}$ and $\mathbf{R}^{21}$ by $\mathbf{C}^1$ and $\mathbf{C}^2$, respectively.

Note, also, that there exists a very close link between VC^3 and VESPA. Although the way we chose $\mathbf{C}^1$ and $\mathbf{C}^2$ above seems to be not very obvious, there is a unique geometric interpretation to it. According to VC^3, as far as the bearing information is concerned, $\mathbf{C}^1$ is equivalent to the autocorrelation matrix of the array, and $\mathbf{C}^2$ is equivalent to the cross-correlation matrix between the array and its virtual copy (which is created by displacing the array by the vector that connects the second and the first sensors).

If the noise component of the signal received by one of the guiding sensor pair elements is independent of the noises at the other sensors, VESPA suppresses the noise regardless of its distribution [6]. In practice, the noise does affect the standard deviations of results obtained from VESPA.

An iterative version of VESPA has also been developed for cases where the source powers have a high dynamic range [11]. Iterative VESPA has the same hardware requirements and assumptions as in VESPA.

Extended VESPA [10]: When there are coherent (or completely correlated) sources, all of the above second- and higher-order statistics methods, except for the WSF method and other MD search-based approaches, fail. For the WSF and other MD methods, however, the array must be calibrated accurately and the computational load is expensive. The coherent signals case arises in practice when there are multipaths. Porat and Friedlander present a modified version of their algorithm to handle the case of coherent signals; however, their method is not practical because it requires selection of a highly redundant subset of fourth-order cumulants that contains $O(N^4)$ elements, and no guidelines exist for its selection and second-, fourth-, sixth-, and eighth-order moments of the data are required. If the array is "uniform linear," coherence can be handled using spatial smoothing as a preprocessor to the usual second- or higher-order [3,39] methods; however, the array aperture is reduced. Extended VESPA can handle coherence and provides increased aperture. Additionally, the array does not have to be completely uniform linear or calibrated; however, a uniform linear subarray is still needed. An example array configuration is shown in Figure 3.3.

Consider a scenario in which there are G statistically independent narrowband sources, $\{u_g(t)\}_{i=1}^G$. These source signals undergo multipath propagation, and each produces p_i coherent wave fronts

$$\{s_{1,1}, \ldots, s_{1,p_1}, \ldots, s_{G,1}, \ldots, s_{G,p_G}\}\left(\sum_{i=1}^{G} p_i = P\right),$$

that impinge on an M element sensor array from directions

$$\{\theta_{1,1}, \ldots, \theta_{1,p_1}, \ldots, \theta_{G,1}, \ldots, \theta_{G,p_G}\},$$

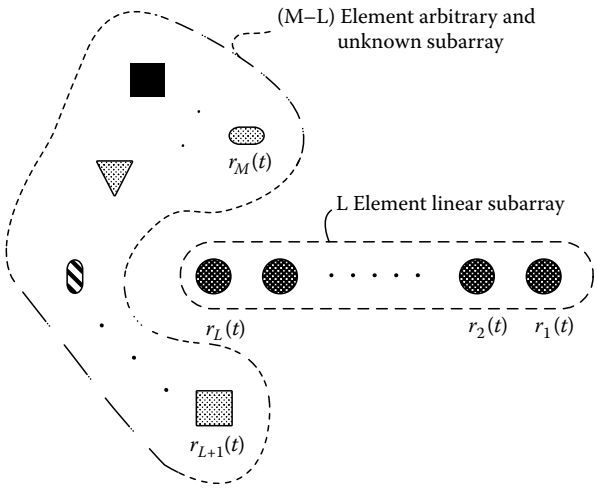

FIGURE 3.3 An example array configuration. There are M sensors, L of which are uniform linearly positioned; $r_1(t)$ and $r_2(t)$ are identical guiding sensors. Linear subarray elements are separated by Δ.

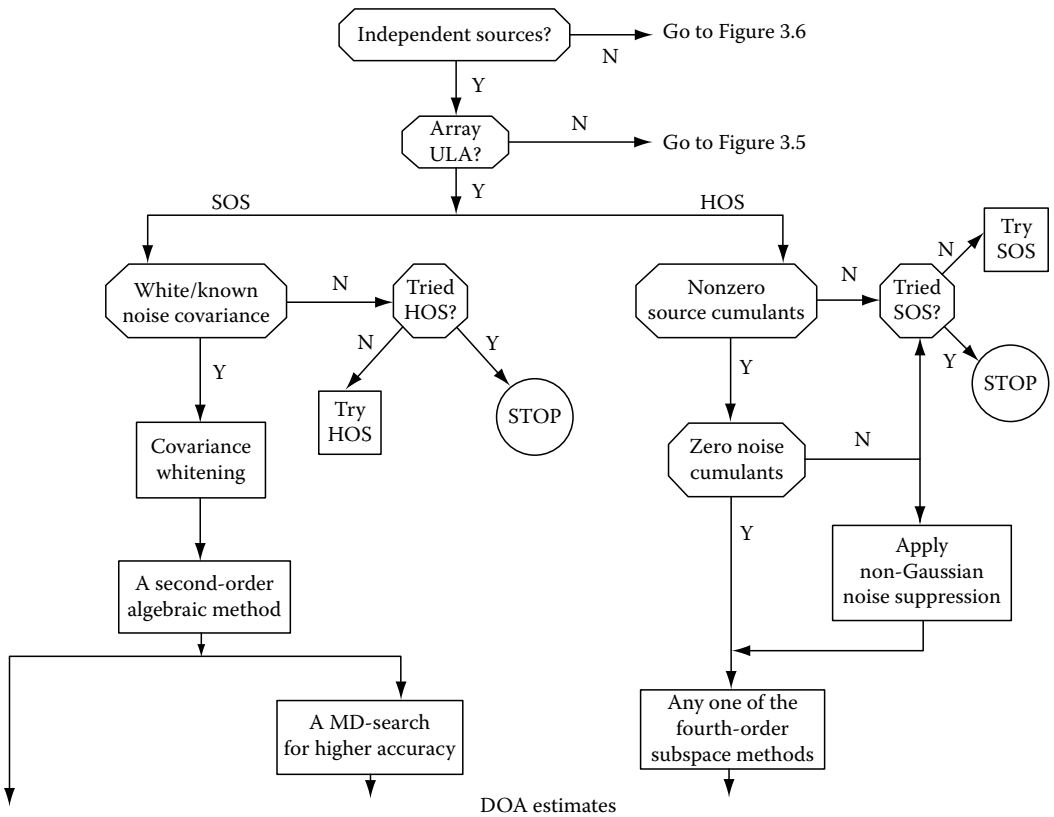

FIGURE 3.4 Second- or higher-order statistics-based subspace DF algorithm. Independent sources and ULA.

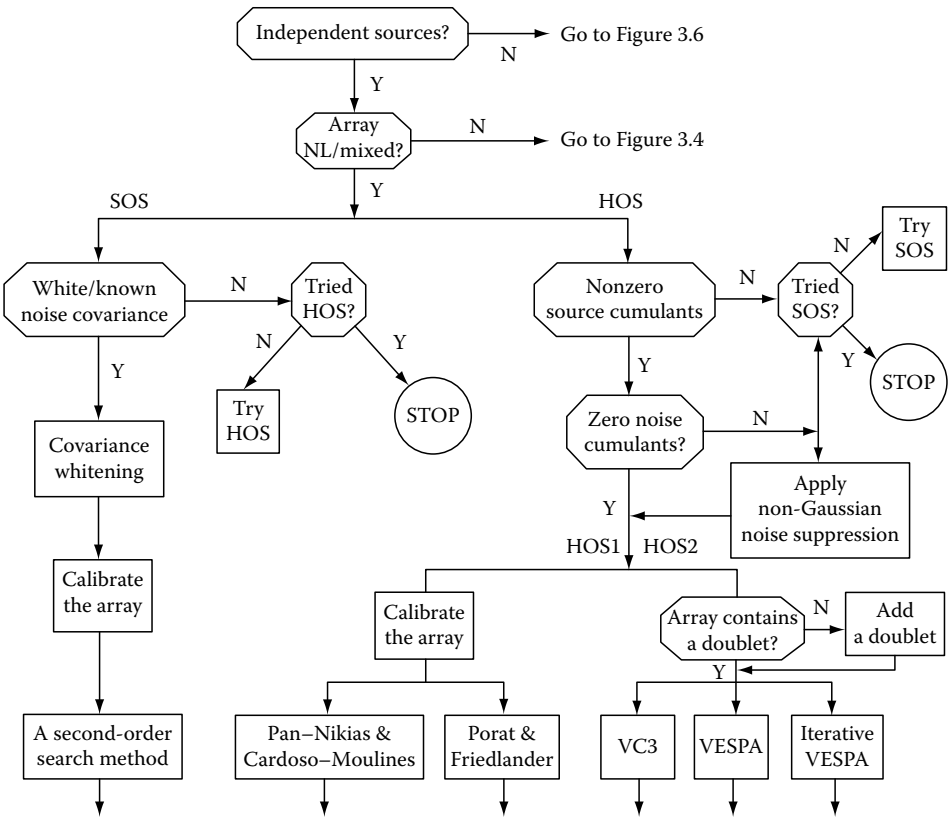

FIGURE 3.5 Second- or higher-order statistics-based subspace DF algorithm. Independent sources and NL/mixed array.

where $\theta_{m,p}$ represents the angle-of-arrival of the wave front $s_{g,p}$ that is the pth coherent signal in the gth group. The collection of p_i coherent wave fronts, which are scaled and delayed replicas of the ith source, are referred to as the ith group. The wave fronts are represented by the P-vector $\mathbf{s}(t)$. The problem is to estimate the DOAs $\{\theta_{1,1}, \ldots, \theta_{1,p_1}, \ldots, \theta_{G,1}, \ldots, \theta_{G,p_G}\}$.

When the multipath delays are insignificant compared to the bit durations of signals, then the signals received from different paths differ by only amplitude and phase shifts, thus the coherence among the received wave fronts can be expressed by the following equation:

$$\mathbf{s}(t) = \begin{bmatrix} \mathbf{s}_1(t) \\ \mathbf{s}_2(t) \\ \vdots \\ \mathbf{s}_G(t) \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{c}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{c}_G \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_G(t) \end{bmatrix} = \mathbf{Q}\mathbf{u}(t), \tag{3.2}$$

where

$\mathbf{s}_i(t)$ is a $p_i \times 1$ signal vector representing the coherent wave fronts from the ith independent source $u_i(t)$

$\mathbf{c}_i$ is a $p_i \times 1$ complex attenuation vector for the ith source ($1 \leq i \leq G$)

$\mathbf{Q}$ is $P \times G$

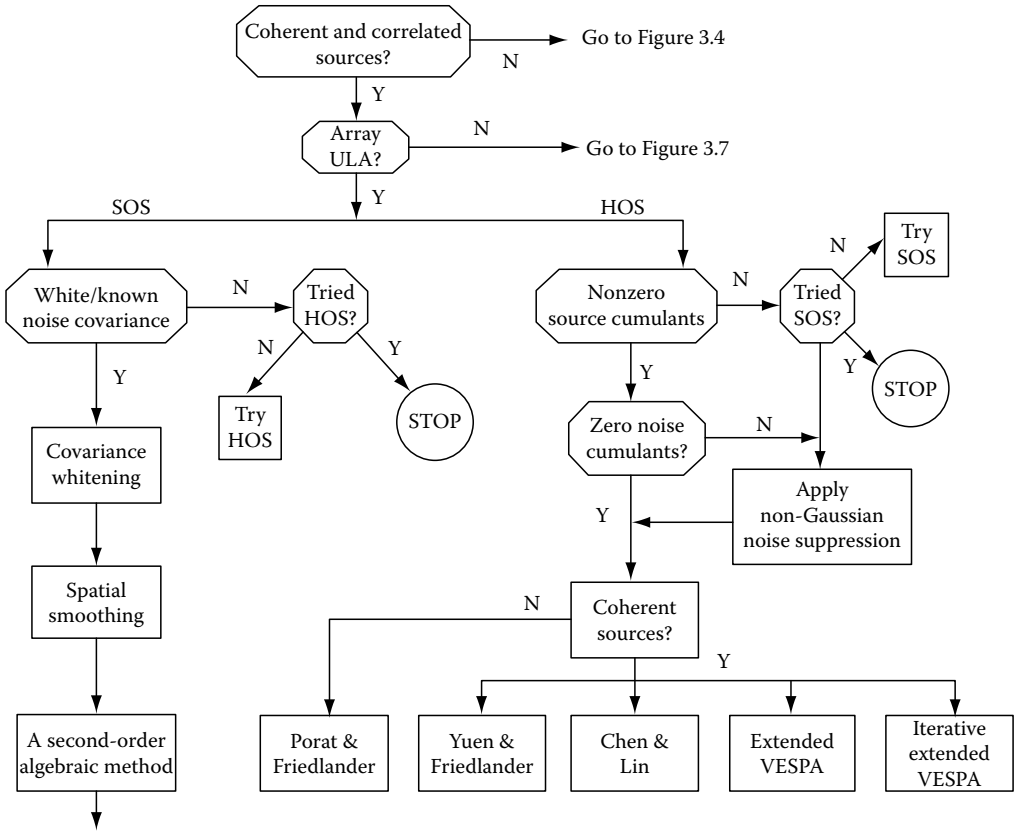

FIGURE 3.6 Second- or higher-order statistics-based subspace DF algorithms. Coherent and correlated sources and ULA.

The elements of $\mathbf{c}_i$ account for the attenuation and phase differences among the multipaths due to different arrival times. The received signal can then be written in terms of the independent sources as follows:

$$\mathbf{r}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(\mathbf{t}) = \mathbf{A}\mathbf{Q}\mathbf{u}(t) + \mathbf{n}(\mathbf{t}) = \mathbf{B}\mathbf{u}(t) + \mathbf{n}(\mathbf{t}), \tag{3.3}$$

where $\mathbf{B} \triangleq \mathbf{AQ}$. The columns of $M \times G$ matrix $\mathbf{B}$ are known as the "generalized steering vectors."

Extended VESPA has three major steps:

Step 1: Use Step (1) of VESPA by choosing $r_1(t)$ and $r_2(t)$ as any two sensor measurements. In this case $\mathbf{C}^1 = \mathbf{BGB}^{\mathrm{H}}$ and $\mathbf{C}^2 = \mathbf{BCGB}^{\mathrm{H}}$, where

$$\mathbf{G} = \mathrm{diag}\left(\gamma_{4,u_1}|b_{11}|^2, \ldots, \gamma_{4,u_G}|b_{1G}|^2\right), \quad \left\{\gamma_{4,u_g}\right\}_{g=1}^{G},$$

$$\mathbf{C} = \mathrm{diag}\left(\frac{b_{21}}{b_{11}}, \ldots, \frac{b_{2G}}{b_{1G}}\right).$$

Due to the coherence, the DOAs cannot be obtained at this step from just $\mathbf{C}^1$ and $\mathbf{C}^2$ because the columns of $\mathbf{B}$ depend on a vector of DOAs (all those within a group). In the independent sources case, the columns of $\mathbf{A}$ depend only on a single DOA. Fortunately, the columns of $\mathbf{B}$ can be solved for as follows:

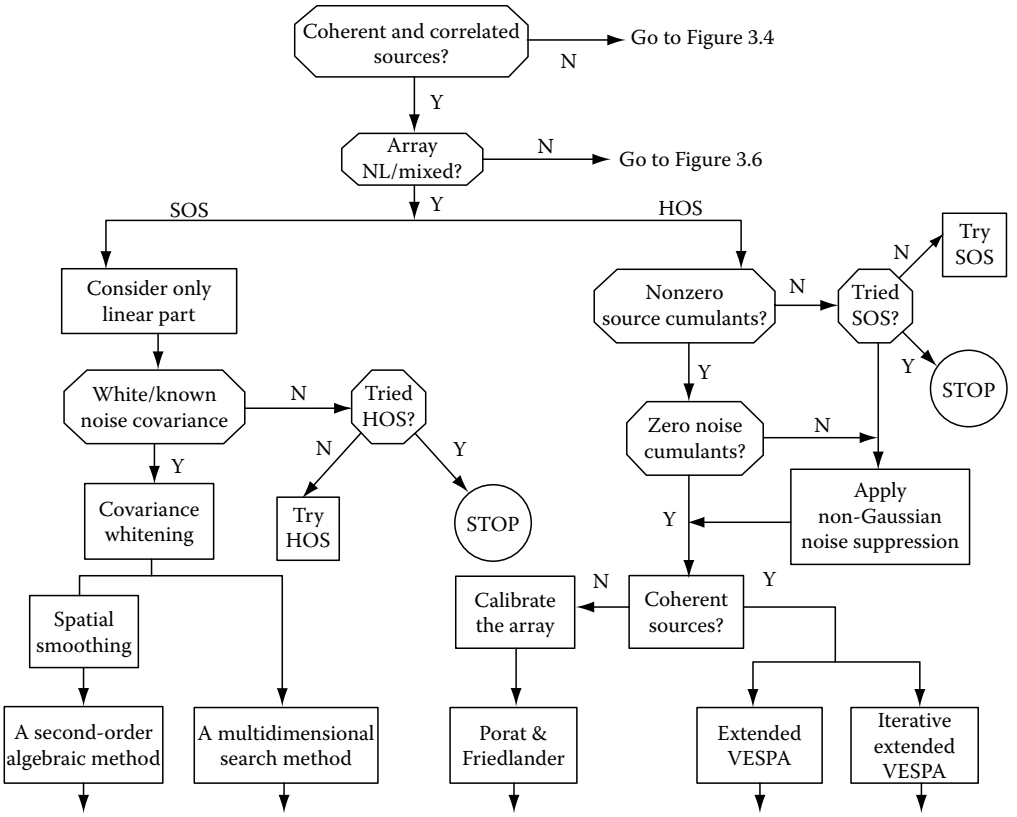

FIGURE 3.7 Second- or higher-order statistics-based subspace DF algorithms. Coherent and correlated sources and NL/mixed array.

(1) follow Steps 2 through 5 of TLS-ESPRIT by replacing $\mathbf{R}^{11}$ and $\mathbf{R}^{21}$ by $\mathbf{C}^1$ and $\mathbf{C}^2$, respectively, and using appropriate matrix dimensions; (2) determine the eigenvectors and eigenvalues of $-\mathbf{F}_x\mathbf{F}_y^{-1}$; let the eigenvector and eigenvalue matrices of $-\mathbf{F}_x\mathbf{F}_y^{-1}$ be $\mathbf{E}$ and $\mathbf{D}$, respectively; and (3) obtain an estimate of $\mathbf{B}$ to within a diagonal matrix, as $\mathbf{B} = (\mathbf{U}_{11}\mathbf{E} + \mathbf{U}_{12}\mathbf{E}\mathbf{D}^{-1})/2$, for use in Step 2.

Step 2: Partition the matrices $\mathbf{B}$ and $\mathbf{A}$ as $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_G]$ and $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_G]$, where the steering vector for the ith group $\mathbf{b}_i$ is $M \times 1$, $\mathbf{A}_i \stackrel{\Delta}{=} [\mathbf{a}(\theta_{i,1}), \ldots, \mathbf{a}(\theta_{i,p_i})]$ is $M \times p_i$, and $\theta_{i,m}$ is the angle-of-arrival of the mth source in the ith coherent group ($1 \leq m \leq p_i$). Using the fact that the ith column of $\mathbf{Q}$ has p_i nonzero elements, express $\mathbf{B}$ as $\mathbf{B} = \mathbf{A}\mathbf{Q} = [\mathbf{A}_1\mathbf{c}_1, \ldots, \mathbf{A}_G\mathbf{c}_G]$; therefore, the ith column of $\mathbf{B}$, $\mathbf{b}_i$ is $\mathbf{b}_i = \mathbf{A}_i\mathbf{c}_i$ where $i = 1, \ldots, G$. Now, the problem of solving for the steering vectors is transformed into the problem of solving for the steering vectors from each coherent group separately. To solve this new problem, each generalized steering vector $\mathbf{b}_i$ can be interpreted as a received signal for an array illuminated by p_i coherent signals having a steering matrix $\mathbf{A}_i$, and covariance matrix $\mathbf{c}_i\mathbf{c}_i^{\mathrm{H}}$. The DOAs could then be solved for by using a second-order-statistics-based high-resolution method such as MUSIC, if the array was calibrated, and the rank of $\mathbf{c}_i\mathbf{c}_i^{\mathrm{H}}$ was p_i; however, the array is not calibrated and rank$\left(\mathbf{c}_i\mathbf{c}_i^{\mathrm{H}}\right) = 1$. The solution is to keep the portion of each $\mathbf{b}_i$ that corresponds to the uniform linear part of the array, $\mathbf{b}_{L,i}$, and to then apply the Section 3.3.3 spatial smoothing technique to a pseudo-covariance matrix $\mathbf{b}_{L,i}\mathbf{b}_{L,i}^{\mathrm{H}}$ for $i = 1, \ldots, G$. Doing this "restores" the rank of $\mathbf{c}_i\mathbf{c}_i^{\mathrm{H}}$ to p_i. In Section 3.3.3, we must replace $\mathbf{r}(t)$ by $\mathbf{b}_{L,i}$ and set $N = 1$.

The conditions on the length of the linear subarray and the parameter S under which the rank of $\mathbf{b}_{S,i}\mathbf{b}_{S,i}^{\mathrm{H}}$ is restored to p_i are [11]: (a) $L \geq 3p_i/2$, which means that the linear subarray must have at least

Second-Order Statistics based Subspace Methods for Direction Finding

Signal Subspace Methods

> *SNR is enhanced effectively by retaining the signal subspace only*

Search Based Methods

> *Select if array is calibrated or response is known analytically*

Correlogram
> *Lower resolution than MV and AR*

Minimum Variance (MV)
> *Narrower mainlobe and smoother sidelobes than conventional beamformers*
> *Higher resolution than Correlogram*
> *Lower resolution than AR*
> *Lower variance than AR*

Autoregressive (AR)
> *Higher resolution than MV and Correlogram*

Subspace Fitting (SF)
> *Weighted SF works regardless of source correlation, and has the same asymptotic properties as the stochastic ML method, i.e., it achieves CRB.*
> *Requires accurate calibration of the manifold and its derivative with respect to arrival angle*

Algebraic Methods

> *Select if the array is ULA or its identical copy exists*
> *Computationally simpler than search-based methods.*

ESPRIT
> *Select if the array has an identical copy*
> *Computationally simple as compared to search-based methods*
> *Sensitive to perturbations in the sensor response and array geometry*
> *LS and TLS versions are best. They have the same asymptotic performance, but TLS converges faster and is better than LS for low SNR and short data lengths*

Toeplitz Approximation Method (TAM)
> *Equivalent to LS-ESPRIT*

GEESE
> *Better than ESPRIT*

Noise Subspace Methods

> *Methods are based on the orthogonality of steering vectors and noise subspace eigenvectors*

Search Based Methods

> *Select if array is calibrated or response is known analytically*

Eigenvector (EV)
> *Produces fewer spurious peaks than MUSIC*
> *Shapes the noise spectrum better than MUSIC*

Pisarenko
> *Performance with short data is poor*

MUSIC
> *Better than MV*
> *Same asymptotic performance as the deterministic ML for uncorrelated sources*

Minimum Norm
> *Select if the array is ULA*
> *Lower SNR threshold than MUSIC for resolution of closely spaced sources*

Method of Direction Estimation (MODE)
> *Consistent for ergodic and stationary signals*

Algebraic Methods

> *Select if the array is ULA*
> *Algebraic versions of EV, Pisarenko, MUSIC, and Minimum Norm are possible*
> *Better resolution than search-based versions*

Root MUSIC
> *Lower SNR threshold than MUSIC for resolution of closely spaced sources*
> *Simple root-finding procedure*

FIGURE 3.8 Pros and cons of all the methods considered.

$3p_{\max}/2$ elements, where $p_{\max}$ is the maximum number of multipaths in anyone of the G groups; and (b) given L and $p_{\max}$, the parameter S must be selected such that $p_{\max} + 1 \leq S \leq L - p_{\max}/2 + 1$.

Step 3: Apply any second-order-statistics-based subspace technique (e.g., root-MUSIC, etc.) to R_i^{fb} $(i = 1, \ldots, G)$ to estimate DOAs of up to $2L/3$ coherent signals in each group.

Note that the matrices $\mathbf{C}$ and $\mathbf{G}$ in $\mathbf{C}^1$ and $\mathbf{C}^2$ are not used; however, if the received signals are independent, choosing $r_1(t)$ and $r_2(t)$ from the linear subarray lets DOA estimates be obtained from $\mathbf{C}$ in Step 1 because, in that case,

$$\mathbf{C} = \text{diag}\left\{ e^{-j2\pi\frac{d}{\lambda}\sin\theta_1}, \ldots, e^{-j2\pi\frac{d}{\lambda}\sin\theta_P} \right\};$$

hence, extended VESPA can also be applied to the case of independent sources.

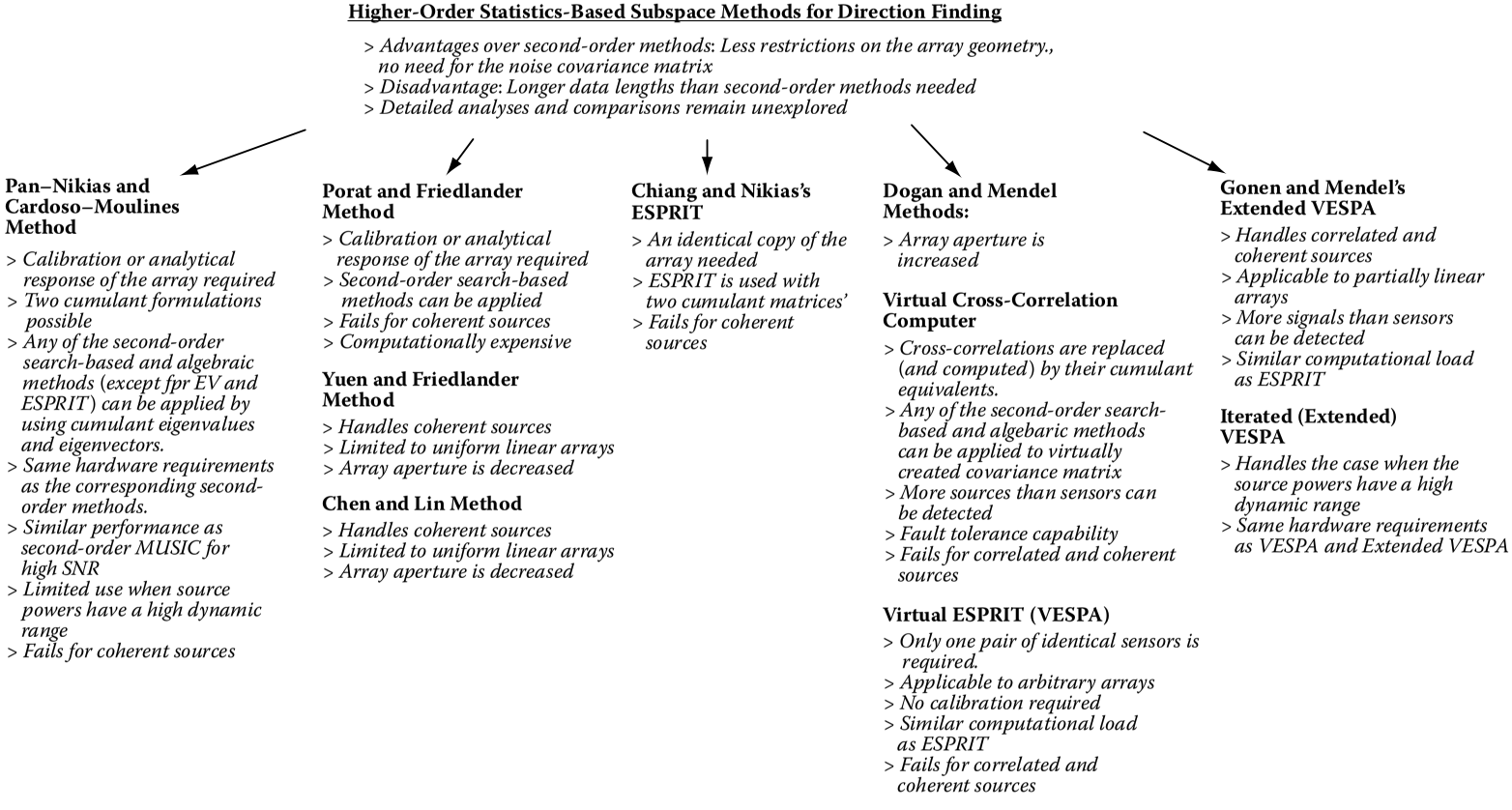

FIGURE 3.9 Pros and cons of all the methods considered.

3.4.1 Discussion

One advantage of using higher-order statistics-based methods over second-order methods is that the covariance matrix of the noise is not needed when the noise is Gaussian. The fact that higher-order statistics have more arguments than covariances leads to more practical algorithms that have less restrictions on the array structure (for instance, the requirement of maintaining identical arrays for ESPRIT is reduced to only maintaining two identical sensors for VESPA). Another advantage is more sources than sensors can be detected, i.e., the array aperture is increased when higher-order statistics are properly applied; or, depending on the array geometry, unreliable sensor measurements can be replaced by using the VC^3 idea. One disadvantage of using higher-order statistics-based methods is that sample estimates of higher-order statistics require longer data lengths than covariances; hence, computational complexity is increased. In their recent study, Cardoso and Moulines [2] present a comparative performance analysis of second- and fourth-order statistics-based MUSIC methods. Their results indicate that dynamic range of the sources may be a factor limiting the performance of the fourth-order statistics-based MUSIC. A comprehensive performance analysis of the above higher-order statistical methods is still lacking; therefore, a detailed comparison of these methods remains as a very important research topic.

3.5 Flowchart Comparison of Subspace-Based Methods

Clearly, there are many subspace-based direction-finding methods. In order to see the forest from the trees, to know when to use a second-order or a higher-order statistics-based method, we present Figures 3.4 through 3.9. These figures provide a comprehensive summary of the existing subspace-based methods for direction finding and constitute guidelines to selection of a proper direction-finding method for a given application.

Note that: Figure 3.4 depicts independent sources and ULA, Figure 3.5 depicts independent sources and NL/mixed array, Figure 3.6 depicts coherent and correlated sources and ULA, and Figure 3.7 depicts coherent and correlated sources and NL/mixed array.

All four figures show two paths: SOS (second-order statistics) and HOS (higher-order statistics). Each path terminates in one or more method boxes, each of which may contain a multitude of methods. Figures 3.8 and 3.9 summarize the pros and cons of all the methods we have considered in this chapter.

Using Figures 3.4 through 3.9, it is possible for a potential user of a subspace-based direction-finding method to decide which method(s) is (are) most likely to give best results for his/her application.

Acknowledgments

The authors would like to thank Profs. A. Paulraj, V.U. Reddy, and M. Kaveh for reviewing the manuscript.

References

1. Capon, J., High-resolution frequency-wavenumber spectral analysis, *Proc. IEEE*, 57(8), 1408–1418, August 1969.
2. Cardoso, J.-F. and Moulines, E., Asymptotic performance analysis of direction-finding algorithms based on fourth-order cumulants, *IEEE Trans. Signal Process.*, 43(1), 214–224, January 1995.
3. Chen, Y.H. and Lin, Y.S., A modified cumulant matrix for DOA estimation, *IEEE Trans. Signal Process.*, 42, 3287–3291, November 1994.
4. Chiang, H.H. and Nikias, C.L., The ESPRIT algorithm with higher-order statistics, in *Proceedings of the Workshop on Higher-Order Spectral Analysis*, Vail, CO, pp. 163–168, June 28–30, 1989.

5. Dogan, M.C. and Mendel, J.M., Applications of cumulants to array processing, Part I: Aperture extension and array calibration, *IEEE Trans. Signal Process*., 43(5), 1200–1216, May 1995.
6. Dogan, M.C. and Mendel, J.M., Applications of cumulants to array processing, Part II: Non-Gaussian noise suppression, *IEEE Trans. Signal Process*., 43(7), 1661–1676, July 1995.
7. Dogan, M.C. and Mendel, J.M., Method and apparatus for signal analysis employing a virtual cross-correlation computer, U.S. Patent No. 5,459,668, October 17, 1995.
8. Ephraim, T., Merhav, N., and Van Trees, H.L., Min-norm interpretations and consistency of MUSIC, MODE and ML, *IEEE Trans. Signal Process*., 43(12), 2937–2941, December 1995.
9. Evans, J.E., Johnson, J.R., and Sun, D.F., High resolution angular spectrum estimation techniques for terrain scattering analysis and angle of arrival estimation, in *Proceedings of the First ASSP Workshop Spectral Estimation*, Communication Research Laboratory, McMaster University, Hamilton, Ontario, Canada, August 1981.
10. Gönen, E., Dogan, M.C., and Mendel, J.M., Applications of cumulants to array processing: Direction finding in coherent signal environment, in *Proceedings of 28th Asilomar Conference on Signals, Systems, and Computers*, Asilomar, CA, pp. 633–637, 1994.
11. Gönen, E., Cumulants and subspace techniques for array signal processing, PhD thesis, University of Southern California, Los Angeles, CA, December 1996.
12. Haykin, S.S., *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1991.
13. Johnson, D.H. and Dudgeon, D.E., *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
14. Kaveh, M. and Barabell, A.J., The statistical performance of the MUSIC and the Minimum-Norm algorithms in resolving plane waves in noise, *IEEE Trans. Acoust. Speech Signal Process*., 34, 331–341, April 1986.
15. Kumaresan, R. and Tufts, D.W., Estimating the angles of arrival multiple plane waves, *IEEE Trans. Aerosp. Electron. Syst*., AES-19, 134–139, January 1983.
16. Kung, S.Y., Lo, C.K., and Foka, R., A Toeplitz approximation approach to coherent source direction finding, *Proceedings of the ICASSP*, Tokyo, Japan, 1986.
17. Li, F. and Vaccaro, R.J., Unified analysis for DOA estimation algorithms in array signal processing, *Signal Process*., 25(2), 147–169, November 1991.
18. Marple, S.L., *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
19. Mendel, J.M., Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications, *Proc. IEEE*, 79(3), 278–305, March 1991.
20. Nikias, C.L. and Petropulu, A.P., *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
21. Ottersten, B., Viberg, M., and Kailath, T., Performance analysis of total least squares ESPRIT algorithm, *IEEE Trans. Signal Process*., 39(5), 1122–1135, May 1991.
22. Pan, R. and Nikias, C.L., Harmonic decomposition methods in cumulant domains, in *Proceedings of the ICASSP'88*, New York, pp. 2356–2359, 1988.
23. Paulraj, A., Roy, R., and Kailath, T., Estimation of signal parameters via rotational invariance techniques-ESPRIT, in *Proceedings of the 19th Asilomar Conference on Signals, Systems, and Computers*, Asilomar, CA, November 1985.
24. Pillai, S.U., *Array Signal Processing*, Springer-Verlag, New York, 1989.
25. Porat, B. and Friedlander, B., Direction finding algorithms based on high-order statistics, *IEEE Trans. Signal Process*., 39(9), 2016–2023, September 1991.
26. Radich, B.M. and Buckley, K., The effect of source number underestimation on MUSIC location estimates, *IEEE Trans. Signal Process*., 42(1), 233–235, January 1994.
27. Rao, D.V.B. and Hari, K.V.S., Performance analysis of Root-MUSIC, *IEEE Trans. Acoust. Speech Signal Process*., ASSP-37, 1939–1949, December 1989.
28. Roy, R.H., ESPRIT-estimation of signal parameters via rotational invariance techniques, PhD dissertation, Stanford University, Stanford, CA, 1987.