THE NEW SPACE ERA

WILEY

Nonlinear Signal Processing

A Statistical Approach

Gonzalo R. Arce

Includes CD-ROM

Nonlinear Signal Processing

This Page Intentionally Left Blank

Nonlinear Signal Processing

A Statistical Approach

Gonzalo R. Arce
University of Delaware
Department of Computer
and Electrical Engineering

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representation or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Arce, Gonzalo R.
Nonlinear signal processing : a statistical approach / Gonzalo R. Arce.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-67624-1 (cloth : acid-free paper)
1. Signal processing—Mathematics. 2. Statistics. I. Title.

TK5102.9.A77 2004
621.382'2—dc22 2004042240

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Catherine, Andrew, Catie, and my beloved parents.

This Page Intentionally Left Blank

Preface

Linear filters today enjoy a rich theoretical framework based on the early and important contributions of Gauss (1795) on Least Squares, Wiener (1949) on optimal filtering, and Widrow (1970) on adaptive filtering. Linear filter theory has consistently provided the foundation upon which linear filters are used in numerous practical applications as detailed in classic treatments including that of Haykin [99], Kailath [110], and Widrow [197]. Nonlinear signal processing, however, offers significant advantages over traditional linear signal processing in applications in which the underlying random processes are nonGaussian in nature, or when the systems acting on the signals of interest are inherently nonlinear. Practice has shown that nonlinear systems and nonGaussian processes emerge in a broad range of applications including imaging, teletraffic, communications, hydrology, geology, and economics. Nonlinear signal processing methods in all of these applications aim at exploiting the system's nonlinearities or the statistical characteristics of the underlying signals to overcome many of the limitations of the traditional practices used in signal processing.

Traditional signal processing enjoys the rich and unified theory of linear systems. Nonlinear signal processing, on the other hand, lacks a unified and universal set of tools for analysis and design. Hundreds of nonlinear signal processing algorithms have been proposed in the literature. Most of the proposed methods, although well tailored for a given application, are not broadly applicable in general. While nonlinear signal processing is a dynamic and rapidly growing field, large classes of nonlinear signal processing algorithms can be grouped and studied in a unified framework. Textbooks on higher-and lower-order statistics [148], polynomial filters [141], neural-networks [100], and mathematical morphology have appeared recently with the common goal of grouping a "self-contained" class of nonlinear signal processing algorithms into a unified framework of study.

This book focuses on unifying the study of a broad and important class of nonlinear signal processing algorithms that emerge from statistical estimation principles, and where the underlying signals are nonGaussian processes. Notably, by concentrating on just two nonGaussian models, a large set of tools is developed that encompasses a large portion of the nonlinear signal processing tools proposed in the literature over the past several decades. In particular, under the generalized Gaussian distribution, signal processing algorithms based on weighted medians and their generalizations are developed. The class of stable distributions is used as the second nonGaussian model from which weighted myriads emerge as the fundamental estimate from which general signal processing tools are developed. Within these two classes of nonlinear signal processing methods, a goal of the book is to develop a unified treatment on optimal and adaptive signal processing algorithms that mirror those of Wiener and Widrow, extensively presented in the linear filtering literature.

The current manuscript has evolved over several years while the author regularly taught a nonlinear signal processing course in the graduate program at the University of Delaware. The book serves an international market and is suitable for advanced undergraduates or graduate students in engineering and the sciences, and practicing engineers and researchers. The book contains many unique features including:

- Numerous problems at the end of each chapter.

- Numerous examples and case studies provided throughout the book in a wide range of applications.

- A set of 60+ MATLAB software m-files allowing the reader to quickly design and apply any of the nonlinear signal processing algorithms described in the book to an application of interest.

- An accompanying MATLAB software guide.

- A companion PowerPoint presentation with more than 500 slides available for instruction.

The chapters in the book are grouped into three parts.

Part I provides the necessary theoretical tools that are used later in text. These include a review of nonGaussian models emphasizing the class of generalized Gaussian distributions and the class of stable distributions. The basic principles of order statistics are covered, which are of essence in the study of weighted medians. Part I closes with a chapter on maximum likelihood and robust estimation principles which are used later in the book as the foundation on which signal processing methods are build upon.

Part II comprises of three chapters focusing on signal processing tools developed under the generalized Gaussian model with an emphasis on the Laplacian model. Weighted medians, L-filters, and several generalizations are studied at length.

Part III encompasses signal processing methods that emerge from parameter estimation within the stable distribution framework.

The chapter sequence is thus assembled in a self-contained and unified framework of study.

This Page Intentionally Left Blank

Acknowledgments

The material in this textbook has benefited greatly from my interaction with many bright students at the University of Delaware. I am particularly indebted to my previous graduate students Juan Gonzalez, Sebastian Hoyos, Sudhakar Kalluri, Yinbo Li, David Griffith, Yeong-Taeg Kim, Edwin Heredia, Alex Flaig, Zhi Zhou, Dan Lau, Karen Bloch, Russ Foster, Russ Hardie, Tim Hall, and Michael McLoughlin. They have all contributed significantly to material throughout the book. I am very grateful to Jan Bacca and Dr. Jose-Luis Paredes for their technical and software contributions. They have generated all of the MATLAB routines included in the book as well as the accompanying software guide. Jan Bacca has provided the much needed electronic publishing support to complete this project.

I am particularly indebted to Dr. Neal C. Gallagher of the University of Central Florida for being a lifelong mentor, supporter, and friend.

It has been a pleasure working with the Non-linear Signal Processing Board: Dr. Hans Burkhardt of the Albert-Ludwigs-University, Freiburg Germany, Dr. Ed Coyle of Purdue University, Dr. Moncef Gabbouj of the Tampere University of Technology, Dr. Murat Kunt of the Swiss Federal Institute of Technology, Dr. Steve Marshall of the University of Strathclyde, Dr. John Mathews of the University of Utah, Dr. Yrjo Neuvo of Nokia, Dr. Ioannis Pitas of the Aristotle University of Thessaloniki, Dr. Jean Serra of the Center of Mathematical Morphology, Dr. Giovanni Sicuranza of the University of Trieste, Dr. Akira Taguchi of the Musashi Institute of Technology, Dr. Anastasios N. Venetsanopoulos of the University of Toronto, and Dr. Pao-Ta Yu of the National Chung Cheng University. Their contribution in the organization of the international workshop series in this field has provided the vigor required for academic excellence.

My interactions with a number of outstanding colleagues has deepened my understanding of nonlinear signal processing. Many of these collaborators have made important contributions to the theory and practice of nonlinear signal processing. I am most grateful to Dr. Ken Barner, Dr. Charles Boncelet, Dr. Xiang Xia, and Dr. Peter Warter all from the University of Delaware, Dr. Jackko Astola, Dr. Karen Egiazarian, Dr. Oli Yli-Harja, Dr I. Tăbus, all from the Tampere University of Technology, Dr. Visa Koivunen of the Helsinki University of Technology, Dr. Saleem Kassam of the University of Pennsylvania, Dr. Sanjit K. Mitra of the University of California, Santa Barbara, Dr. David Munson of the University of Michigan, Dr. Herbert David of Iowa State University, Dr. Kotroupolus of the Universtiy of Thessaloniki, Dr. Yrjo Neuvo of Nokia, Dr. Alan Bovik and Dr. Ilya Shmulevich, both of the University of Texas, Dr. Francesco Palmieri of the University of Naples, Dr. Patrick Fitch of the Lawrence Livermore National Laboratories, Dr. Thomas Nodes of TRW, Dr. Brint Cooper of Johns Hopkins University, Dr. Petros Maragos of the University of Athens, and Dr. Y. H. Lee of KAIST University.

I would like to express my appreciation for the research support I received from the National Science Foundation and the Army Research laboratories, under the Federated Laboratories and Collaborative Technical Alliance programs, for the many years of research that led to this textbook. I am particularly grateful to Dr. John Cozzens and Dr. Taieb Znati, both from NSF, and Dr. Brian Sadler, Dr. Ananthram Swami, and Jay Gowens, all from ARL. I am also grateful to the Charles Black Evans Endowment that supports my current Distinguished Professor appointment at the University of Delaware.

I would like to thank my publisher George Telecki and the staff at Wiley for their dedicated work during this project and Heather King for establishing the first link to Wiley.

G. R. A.

Contents

Part I Statistical Foundations

This Page Intentionally Left Blank

Acronyms

ADSL	Asymmetric digital suscriber line
BIBO	Bounded-input bounded-output
BR	Barrodale and Roberts' (algorithm)
CMA	Constant modulus algorithm
CWM	Center-weighted median
CWMY	Center-weighted myriad
DWMTM	Double window modified Trimmed mean
DWD	Discrete Wigner distribution
FIR	Finite impulse response
FLOS	Fractional lower-order statistics
FLOM	Fractiona lower-order moments
HOS	higher-order statistics
i.i.d	Independent and identically distributed
IIR	Infinite impulse response
LCWM	Linear combination of weighted medians
LS	Least squares
LAD	Least absolute deviation

LLS	Logarithmic least squares
LMS	Least mean square
LMA	Least mean absolute
LP	Linearity parameter
MSE	Mean square error
ML	Maximum likelihood
MAE	Mean absolute error
MTM	Modified trimmed mean
PAM	Phase amplitude modulation
pdf	Portable document format
PLL	Phase lock loop
PSNR	Peak signal-to-noise ratio
PBF	Positive boolean function
RTT	Round trip time
SαS	Symmetric α-stable
SSP	Sample selection probabilities
TCP/IP	Internet transfer protocol
TD	Threshold Decomposition
WM	Weighted median
WMM	Weighted multichannel median
WD	Wigner distribution
ZOS	Zero-order statistics

1

Introduction

Signal processing is a discipline embodying a large set of methods for the representation, analysis, transmission, and restoration of information-bearing signals from various sources. As such, signal processing revolves around the mathematical manipulation of signals. Perhaps the most fundamental form of signal manipulation is that of filtering, which describes a rule or procedure for processing a signal with the goal of separating or attenuating a desired component of an observed signal from either noise, interference, or simply from other components of the same signal. In many applications, such as communications, we may wish to remove noise or interference from the received signal. If the received signal was in some fashion distorted by the channel, one of the objectives of the receiver is to compensate for these disturbances. Digital picture processing is another application where we may wish to enhance or extract certain image features of interest. Perhaps image edges or regions of the image composed of a particular texture are most useful to the user. It can be seen that in all of these examples, the signal processing task calls for separating a desired component of the observed waveform from any noise, interference, or undesired component. This segregation is often done in frequency, but that is only one possibility. Filtering can thus be considered as a system with arbitrary input and output signals, and as such the filtering problem is found in a wide range of disciplines including economics, engineering, and biology.

A classic filtering example, depicted in Figure 1.1, is that of bandpass filtering a frequency rich chirp signal. The frequency components of the chirp within a selected band can be extracted through a number of linear filtering methods. Figure 1.1*b* shows the filtered chirp when a linear 120-tap finite impulse response (FIR) filter is used. This figure clearly shows that linear methods in signal processing can indeed

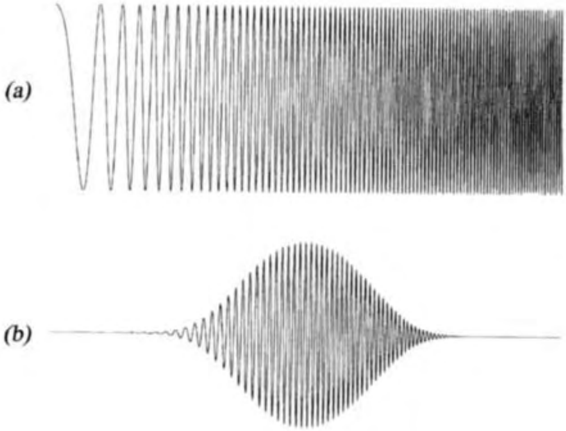


Figure 1.1 Frequency selective filtering: (*a*) chirp signal, (*b*) linear FIR filter output.

be markedly effective. In fact, linear signal processing enjoys the rich theory of linear systems, and in many applications linear signal processing algorithms prove to be optimal. Most importantly, linear filters are inherently simple to implement, perhaps the dominant reason for their widespread use.

Although linear filters will continue to play an important role in signal processing, nonlinear filters are emerging as viable alternative solutions. The major forces that motivate the implementation of nonlinear signal-processing algorithms are the growth of increasingly challenging applications and the development of more powerful computers. Emerging multimedia and communications applications are becoming significantly more complex. Consequently, they require the use of increasingly sophisticated signal-processing algorithms. At the same time, the ongoing advances of computers and digital signal processors, in terms of speed, size, and cost, makes the implementation of sophisticated algorithms practical and cost effective.

Why Nonlinear Signal Processing? Nonlinear signal processing offers advantages in applications in which the underlying random processes are nonGaussian. Practice has shown that nonGaussian processes do emerge in a broad array of applications, including wireless communications, teletraffic, hydrology, geology, economics, and imaging. The common element in these applications, and many others, is that the underlying processes of interest tend to produce more large-magnitude (outlier or impulsive) observations than those that would be predicted by a Gaussian model. That is, the underlying signal density functions have tails that decay at rates lower than the tails of a Gaussian distribution. As a result, linear methods which obey the superposition principle suffer from serious degradation upon the arrival of samples corrupted with high-amplitude noise. Nonlinear methods, on the other hand, exploit the statistical characteristics of the noise to overcome many of the limitations of the traditional practices in signal processing.

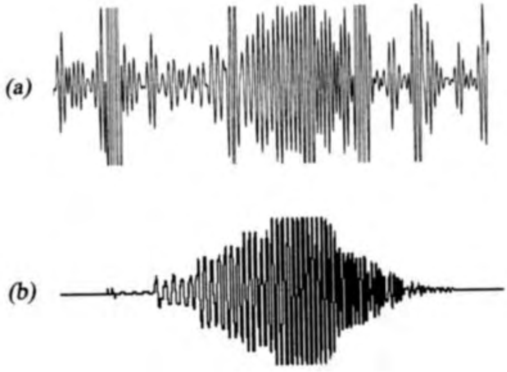


Figure 1.2 Frequency selective filtering in nonGaussian noise: (a) linear FIR filter output, (b) nonlinear filter.

To illustrate the above, consider again the classic bandpass filtering example. This time, however, the chirp signal under analysis has been degraded by nonGaussian noise during the signal acquisition stage. Due to the nonGaussian noise, the linear FIR filter output is severely degraded as depicted in Figure 1.2a. The advantages of an equivalent nonlinear filter are illustrated in Figure 1.2b where the frequency components of the chirp within the selected band have been extracted, and the ringing artifacts and the noise have been suppressed[1].

Internet traffic provides another example of signals arising in practice that are best modeled by nonGaussian models for which nonlinear signal processing offer advantages. Figure 1.3 depicts several round trip time delay series, each of which measures the time that a TCP/IP packet takes to travel between two network hosts. An RTT measures the time difference between the time when a packet is sent and the time when an acknowledgment comes back to the sender. RTTs are important in retransmission transport protocols used by TCP/IP where reliability of communications is accomplished through packet reception acknowledgments, and, when necessary, packet retransmissions. In the TCP/IP protocol, the retransmission of packets is based on the prediction of future RTTs. Figure 1.3 depicts the nonstationary characteristics of RTT processes as their mean varies dramatically with the network load. These processes are also nonGaussian indicating that nonlinear prediction of RTTs can lead to more efficient communication protocols.

Internet traffic exhibits nonGaussian statistics, not only on the RTT delay data mechanisms, but also on the data throughput. For example, the traffic data shown in Figure 1.4 corresponds to actual Gigabit (1000 Mb/s) Ethernet traffic measured on a web server of the ECE Department at the University of Delaware. It was measured using the TCPDUMP program, which is part of the Sun Solaris operating system. To

[1] The example uses a weighted median filter that is developed in later sections.

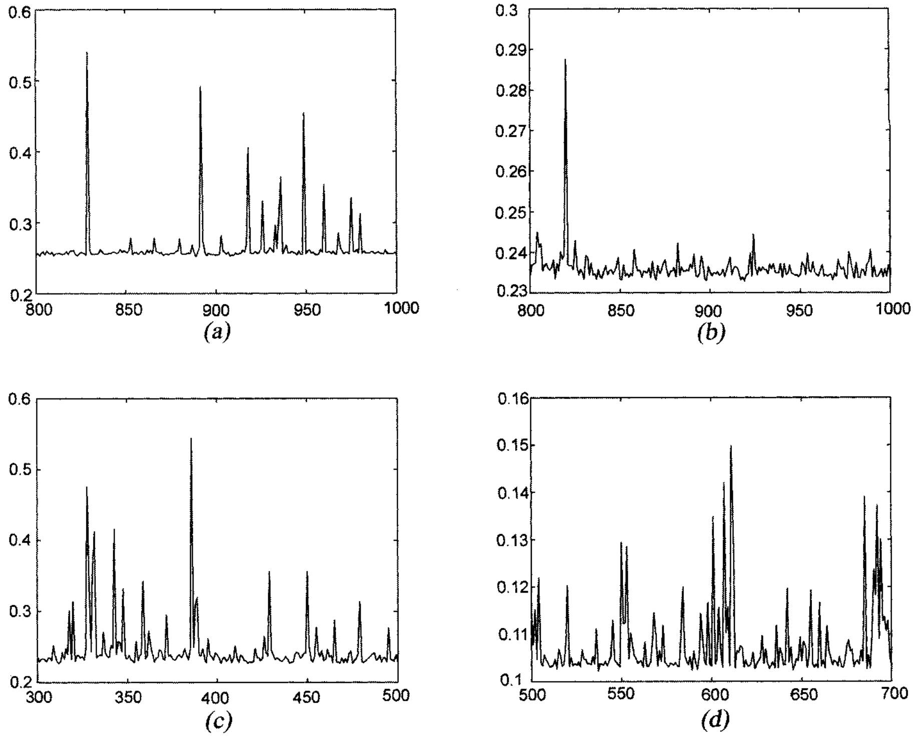


Figure 1.3 RTT time series measured in seconds between a host at the University of Delaware and hosts in (*a*) Australia (12:18 AM - 3:53 AM); (*b*) Sydney, Australia (12:30 AM - 4:03 AM); (*c*) Japan (2:52 PM - 6:33 PM); (*d*) London, UK (10:00 AM - 1:35 PM). All plots shown in 1 minute interval samples.

generate this trace, all packets coming to the server were captured and time-stamped during several hours. The figure considers byte counts (size of the transferred data) measured on 10ms intervals, which is shown in the top plot of Figure 1.4. The overall length of the recordings is approximately four hours (precisely 14,000s). The other plots in Figure 1.4 represent the "aggregated" data obtained by averaging the data counts on increasing time intervals. The notable fact in Figure 1.4 is that the aggregation does not smooth out the data. The aggregated traffic still appears bursty even in the bottom plot despite the fact that each point in it is the average of one thousand successive values of the series shown in the top plot of Figure 1.4. Similar behavior in data traffic has been observed in numerous experimental setups, including Cappé et al. (2002) [42], Beran et al. (1995) [31], Leland et al. (1994) [127], and Paxson and Floyd (1995) [159].

Another example is found in high-speed data links over telephone wires, such as Asymmetric Digital Subscriber Lines (ADSL), where noise in the communications channel exhibits impulsive characteristics. In these systems, telephone twisted pairs

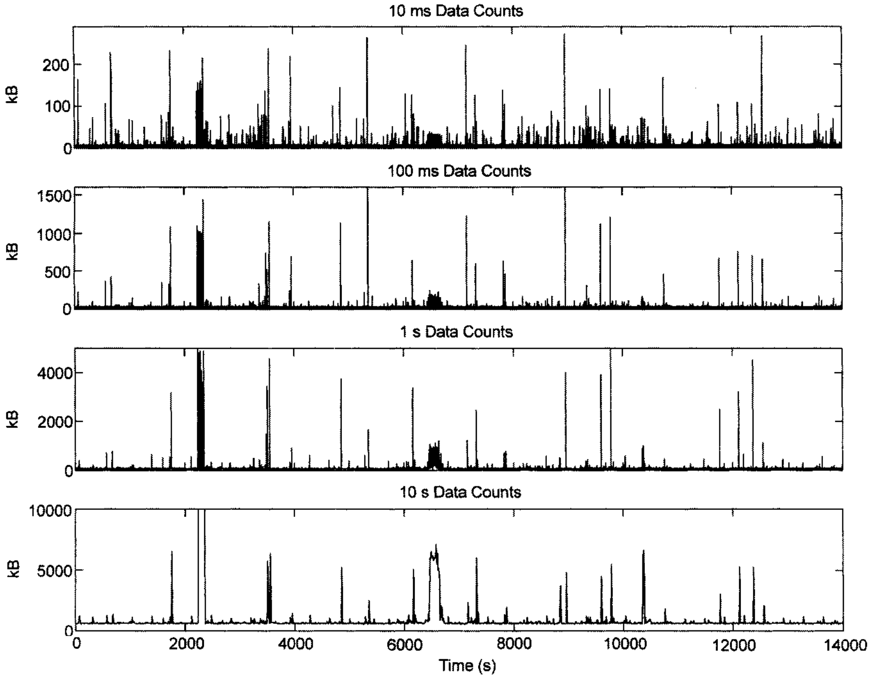


Figure 1.4 Byte counts measured over 14,000 seconds in a web server of the ECE Department at the University of Delaware viewed through different aggregation intervals: from top to bottom, 10ms, 100ms 1s, 10s.

are unshielded, and are thus susceptible to large electromagnetic interference. Potential sources of impulsive interference include light switching and home appliances, as well as natural weather phenomena. Severe interference is also generated by cross talk among multiple twisted pairs making up a telephone cable. The interference is inherently impulsive and nonstationary leading to poor service reliability. The impact of impulsive noise on ADSL systems depends on the impulse energy, duration, interarrival time, and spectral characteristics. Isolated impulses can reach magnitudes significantly larger than either additive white noise or crosstalk interference. A number of models to characterize ADSL interference have been proposed [139]. Current ADSL systems are designed conservatively under the assumption of a worst-case scenario due to severe nonstationary and nonGaussian channel interference [204]. Figure 1.5 shows three ADSL noise signals measured at a customer's premise. These signals exhibit a wide range of spectral characteristics, burstiness, and levels of impulsiveness. In addition to channel coding, linear filtering is used to combat ADSL channel interference [204]. Figure 1.5*a–c* depicts the use of linear and nonlinear filtering. These figures depict the improvement attained by nonlinear filtering in removing the noise and interference.

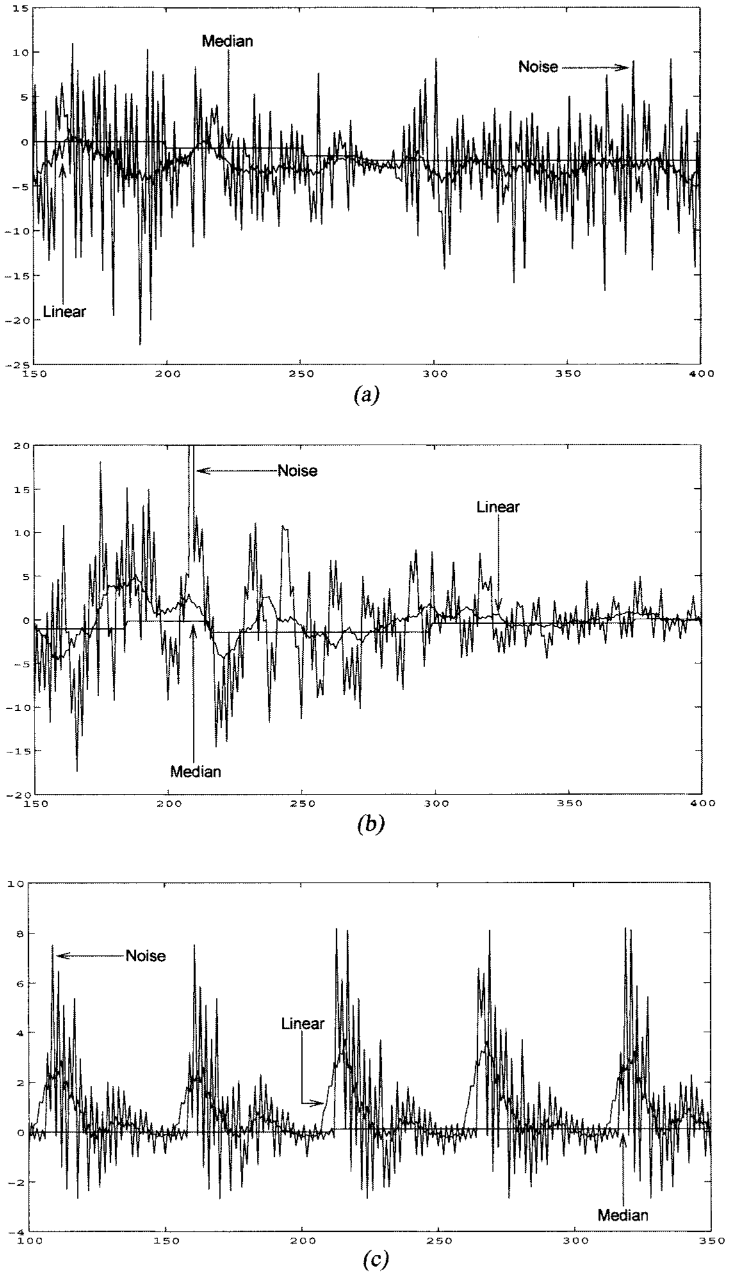


Figure 1.5 (*a–c*) Different noise and interference characteristics in ADSL lines. A linear and a nonlinear filter (recursive median filter) are used to overcome the channel limitations, both with the same window size (adapted from [204]).

The last example (Fig. 1.6), visually illustrates the advantages of nonlinear signal processing. This figure depicts an enlarged section of an image which has been JPEG compressed for storage in a Web site. Since compression reduces and often eliminates the high frequency components, compressed images contain edge artifacts and tend to look blurred. As a result, images found on the Internet are often sharpened. Figure 1.6*b* shows the output of a traditional sharpening algorithm equipped with linear FIR filters. The amplification of the compression artifacts are clearly seen. Figure 1.6*c* depicts the sharpening output when nonlinear filters are used. Nonlinear sharpeners avoid noise and artifact amplification and are as effective as linear sharpeners in highlighting the signal edges.

The examples above suggest that significant improvements in performance can be achieved by nonlinear methods of signal processing. Unlike linear signal processing, however, nonlinear signal processing lacks a unified and universal set of tools for analysis and design. Hundreds of nonlinear signal processing algorithms have been proposed [21, 160]. While nonlinear signal processing is a dynamic, rapidly growing field, a large class of nonlinear signal algorithms can be studied in a unified framework. Since signal processing focuses on the analysis and transformation of signals, nonlinear filtering emerges as the fundamental building block of nonlinear signal processing. This book develops the fundamental signal-processing tools that arise when considering the filtering of nonGaussian, rather than Gaussian, random processes. By concentrating on just two nonGaussian models, a large set of tools is developed that notably encompass a significant portion of the nonlinear signal-processing tools proposed in the literature over the past several decades.

1.1 NONGAUSSIAN RANDOM PROCESSES

In statistical signal processing, signals are modeled as random processes and many signal-processing tasks reduce to the proper statistical analysis of the observed signals. Selecting the appropriate model for the application at hand is of fundamental importance. The model, in turn, determines the signal processing approach. Classical linear signal-processing methods rely on the popular Gaussian assumption. The Gaussian model appears naturally in many applications as a result of the Central Limit Theorem first proved by De Moivre (1733) [69].

THEOREM 1.1 (CENTRAL LIMIT THEOREM) *Let $X_1, X_2, \ldots,$ be a sequence of i.i.d. random variables with zero mean and variance σ^2. Then as $N \to \infty$, the normalized sum*

$$S_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \tag{1.1}$$

converges almost surely to a zero-mean Gaussian variable with the same variance as X_i.

Conceptually, the central limit theorem explains the Gaussian nature of processes generated from the superposition of many small and independent effects. For ex-

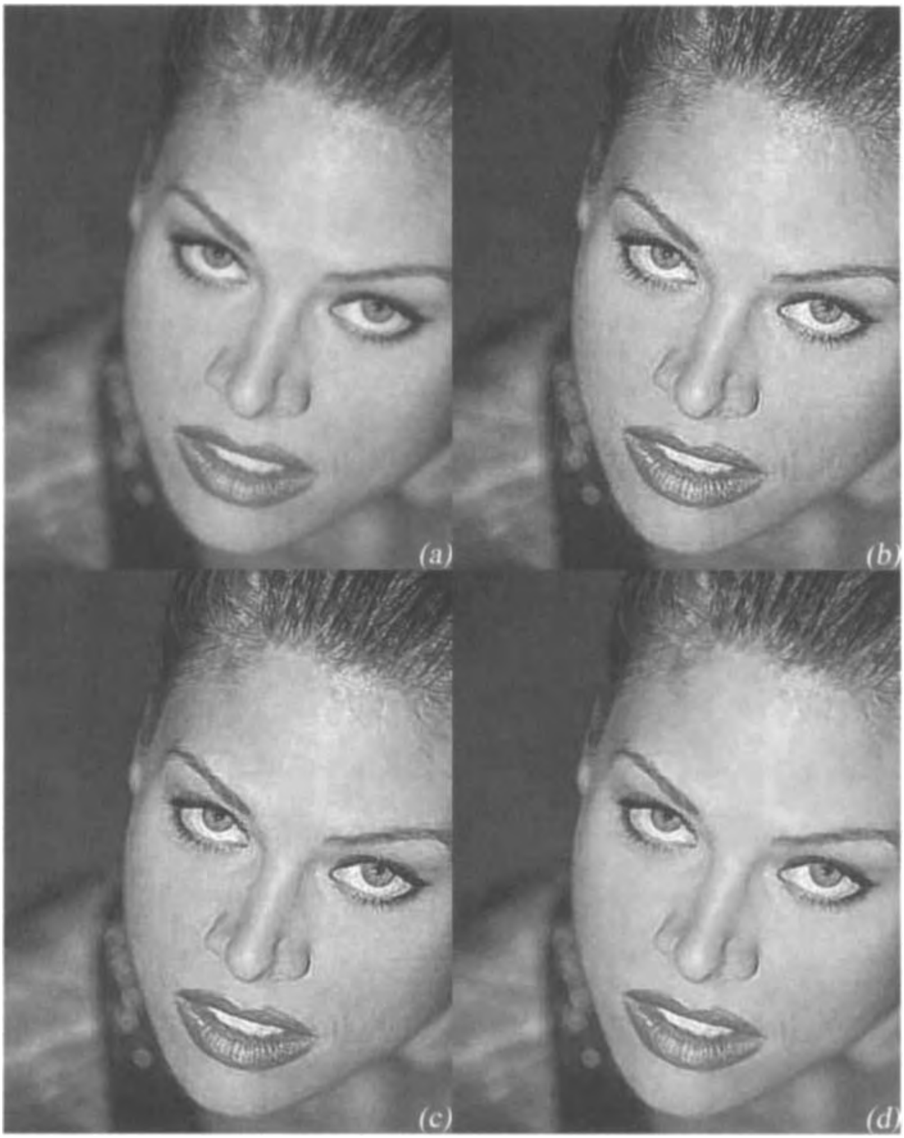


Figure 1.6 (*a*) Enlarged section of a JPEG compressed image, (*b*) output of unsharp masking using FIR filters, (*c*) and (*d*) outputs of median sharpeners.

ample, thermal noise generated as the superposition of a large number of random independent interactions at the molecular level. The Central Limit Theorem theoretically justifies the appearance of Gaussian statistics in real life.

However, in a wide range of applications, the Gaussian model does not produce a good fit which, at first, may seem to contradict the principles behind the Central Limit Theorem. A careful revision of the conditions of the Central Limit Theorem indicates that, in order for this theorem to be valid, the variance of the superimposed random variables must be finite. If the random variables possess infinite variance, it can be shown that the series in the Central Limit Theorem converges to a non-Gaussian impulsive variable [65, 207]. This important generalization of the Central Limit Theorem explains the apparent contradictions of its "traditional" version, as well as the presence of non-Gaussian, infinite variance processes, in practical problems. In the same way as the Gaussian model owes most of its strength to the Central Limit Theorem, the Generalized Central Limit Theorem constitutes a strong theoretical argument to the development of models that capture the impulsive nature of these signals, and of signal processing tools that are adequate in these nonGaussian environments.

Perhaps the simplest approach to address the effects of nonGaussian signals is to detect outliers that may be present in the data, reject these heuristically, and subsequently use classical signal-processing algorithms. This approach, however, has many disadvantages. First, the detection of outliers is not simple, particularly when these are bundled together. Second, the efficiency of these methods is not optimal and is generally difficult to measure since the methods are based on heuristics.

The approach followed in this book is that of exploiting the rich theories of robust statistics and non-Gaussian stochastic processes, such that a link is established between them leading to signal processing with solid theoretical foundations. This book considers two model families that encompass a large class of random processes. These models described by their distributions allow the rate of tail decay to be varied: the *generalized Gaussian* distribution and the class of *stable* distributions. The tail of a distribution can be measured by the mass of the tail, or *order*, defined as $P_r(X > x)$ as $x \rightarrow \infty$. Both distribution families are general in that they encompass a wide array of distributions with different tail characteristics. Additionally, both the generalized Gaussian and stable distributions contain important special cases that lead directly to classes of nonlinear filters that are tractable and optimal for signals with heavy tail distributions.

1.1.1 Generalized Gaussian Distributions and Weighted Medians

One approach to modeling the presence of outliers is to start with the Gaussian distribution and allow the exponential rate of tail decay to be a free parameter. This results directly in the generalized Gaussian density function. Of special interest is the case of first order exponential decay, which yields the double exponential, or Laplacian, distribution. Optimal estimators for the generalized Gaussian distribution take on a particularly simple realization in the Laplacian case. It turns out that weighted median filters are optimal for samples obeying Laplacian statistics, much like linear filters are optimal for Gaussian processes. In general, weighted median filters are more efficient than linear filters in impulsive environments, which can be directly attributed to the heavy tailed characteristic of the Laplacian distribution. Part II of the book uncovers signal processing methods using median-like operations, or order statistics.

1.1.2 Stable Distributions and Weighted Myriads

Although the class of generalized Gaussian distributions includes a spectrum of impulsive processes, these are all of exponential tails. It turns out that a wide variety of processes exhibit more impulsive statistics that are characterized with algebraic tailed distributions. These impulsive processes found in signal processing applications arise as the superposition of many small independent effects. For example, radar clutter is the sum of many signal reflections from an irregular surface; the transmitters in a multiuser communication system generate relatively small independent signals, the sum of which represents the ensemble at a user's receiver; rotating electric machinery generates many impulses caused by contact between distinct parts of the machine; and standard atmospheric noise is known to be the superposition of many electrical discharges caused by lightning activity around the Earth. The theoretical justification for using stable distribution models lies in the Generalized Central Limit Theorem which includes the well known "traditional" Central Limit Theorem as a special case. Informally:

> *A random variable X is stable if it can be the limit of a normalized sum of i.i.d. random variables.*

The generalized theorem states that if the sum of i.i.d. random variables with or without finite variance converges to a distribution, the limit distribution must belong to the family of stable laws [149, 207]. Thus, nonGaussian processes can emerge in practical applications as sums of random variables in the same way as Gaussian processes.

Stable distributions include two special cases of note: the standard Gaussian distribution and the Cauchy distribution. The Cauchy distribution is particularly important as its tails decay algebraically. Thus, the Cauchy distribution can be used to model very impulsive processes. It turns out that for a wide range of stable-distributed signals, the so-called weighted myriad filters are optimal. Thus, weighted myriad filters emerging from the stable model are the counterparts to linear and median filters related to the Gaussian and Laplacian environments, respectively. Part III of the book develops signal-processing methods derived from stable models.

1.2 STATISTICAL FOUNDATIONS

Estimation theory is a branch of statistics concerned with the problem of deriving information about the properties of random processes from a set of observed samples. As such, estimation theory lies at the heart of statistical signal processing. Given an observation waveform $\{X(n)\}$, one goal is to extract information that is embedded within the observed signal. It turns out that the embedded information can often be modeled parametrically. That is, some parameter β of the signal represents the information of interest. This parameter may be the local mean, the variance, the local range, or some other parameter associated with the received waveform. Of course, finding a good parametric model is critical.

Location Estimation Because observed signals are inherently random, these are described by a probability density function (pdf), $f(x_1, x_2, \ldots, x_N)$. The pdf may be parameterized by an unknown parameter β. The parameter β thus defines a class of pdfs where each member is defined by a particular value of β. As an example, if our signal consists of a single point ($N = 1$) and β is the mean, the pdf of the data under the Gaussian model is

$$f(x_1; \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(x_1 - \beta)^2\right] \tag{1.2}$$

which is shown in Figure 1.7 for various values of β. Since the value of β affects the probability of X_1, intuitively we should be able to infer the value of β from the observed value of X_1. For example, if the observed value of X_1 is a large positive number, the parameter β is more likely to be equal to β_1 than to β_2 in Figure 1.7. Notice that β determines the location of the pdf. As such, β is referred to as the *location parameter.* Rules that infer the value of β from sample realizations of the data are known as *location estimators.* Although a number of parameters can be associated with a set of data, *location* is a parameter that plays a key role in the design of filtering algorithms. The filtering structures to be defined in later chapters have their roots in location estimation.

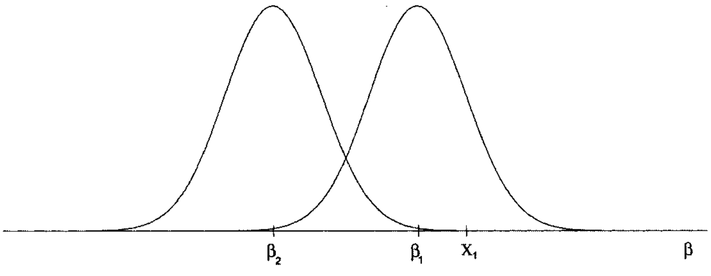


Figure 1.7 Estimation of parameter β based on the observation X_1.

Running Smoothers Location estimation and filtering are intimately related. The *running mean* is the simplest form of filtering and is most useful in illustrating this relationship. Given the data sequence $\{\ldots, X(n-1), X(n), X(n+1), \ldots\}$, the running mean is defined as

$$Y(n) = \mathrm{MEAN}(X(n-N), X(n-N+1), \ldots, X(n+N)). \tag{1.3}$$

At a given point n, the output is the average of the samples within a window centered at n. The output at $n+1$ is the average of the samples within the window centered at $n+1$, and so on. Thus, at each point n, the running mean computes a location estimate, namely the sample mean. If the underlying signals are not Gaussian, it would be reasonable to replace the mean by a more appropriate location estimator. Tukey (1974) [189], for instance, introduced the running median as a robust alternative to the running mean.

Although running smoothers are effective in removing noise, more powerful signal processing is needed in general to adequately address the tasks at hand. To this end, the statistical foundation provided by running smoothers can be extended to define optimal filtering structures.

1.3 THE FILTERING PROBLEM

Filtering constitutes a system with arbitrary input and output signals, and consequently the filtering problem is found in a wide range of disciplines. Although filtering theory encompasses continuous-time as well as discrete-time signals, the availability of digital computer processors is causing discrete-time signal representation to become the preferred method of analysis and implementation. In this book, we thus consider signals as being defined at discrete moments in time where we assume that the sampling interval is fixed and small enough to satisfy the Nyquist sampling criterion.

Denote a random sequence as $\{X\}$ and let $\mathbf{X}(n)$ be a N-long element, real valued observation vector

$$\begin{aligned}\mathbf{X}(n) &= [X(n),\ X(n-1), \ldots, X(n-N+1)]^T \\ &= [X_1(n),\ X_2(n), \ldots,\ X_N(n)]^T\end{aligned} \tag{1.4}$$

where $X_i(n) = X(n-i+1)$ and where T denotes the transposition operator. R denotes the real line. Further, assume that the observation vector $\mathbf{X}(n)$ is statistically related to some desired signal denoted as $D(n)$. The filtering problem is then formulated in terms of joint process estimation as shown in Figure 1.8. The observed vector, $\mathbf{X}(n)$, is formed by the elements of a shifting window, the output of the filter is the estimate $\hat{D}(n)$ of a desired signal $D(n)$. The optimal filtering problem thus reduces to minimizing the cost function associated with the error $e(n)$ under a given criterion, such as the mean square error (MSE).

Under Gaussian statistics, the estimation framework becomes linear and the filter structure reduces to that of FIR linear filters. The linear filter output is defined as

$$Y(n) = \mathrm{MEAN}(W_1 \cdot X_1(n), W_2 \cdot X_2(n), \ldots, W_N \cdot X_N(n)), \tag{1.5}$$

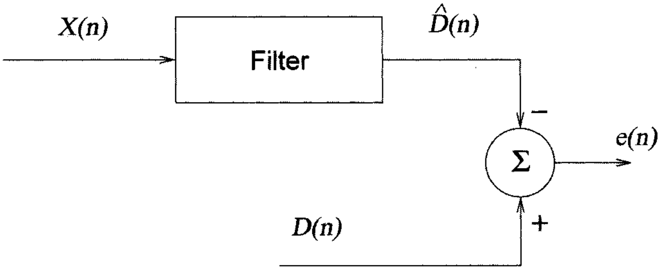


Figure 1.8 Filtering as a joint process estimation

where the W_i are real-valued weights assigned to each input sample.

Under the Laplacian model, it will be shown that the median becomes the estimate of choice and weighted medians become the filtering structure. The output of a weighted median is defined as

$$Y(n) = \text{MEDIAN}(W_1 \diamond X_1(n), W_2 \diamond X_2(n), \ldots, W_N \diamond X_N(n)), \tag{1.6}$$

where the operation $W_i \diamond X_i(n)$ replicates the sample $X_i(n)$, W_i times. Weighting in median filters thus takes on a very different meaning than traditional weighting in linear filters.

For stable processes, it will be derived shortly that the weighted myriad filter emerges as the ideal structure. In this case the filter output is defined as

$$Y(n) = \text{MYRIAD}\left(K;\ W_1 \circ X_1, W_2 \circ X_2, \ldots, W_N \circ X_N\right), \tag{1.7}$$

where $W_i \circ X_i(n)$ represents a nonlinear weighting operation to be described later, and K in (1.7) is a free tunable parameter that will play an important role in weighted myriad filtering. It is the flexibility provided by K that makes the myriad filter a more powerful filtering framework than either the linear FIR or the weighted median filter frameworks.

1.3.1 Moment Theory

Historically, signal processing has relied on second-order moments, as these are intimately related to Gaussian models. The first-order moment

$$\mu_X = E\{X(n)\} \tag{1.8}$$

and the second-order moment characterization provided by the *autocorrelation* of stationary processes

$$R_X(k) = E\left\{X(n)X(n+k)\right\} \tag{1.9}$$

are deeply etched into traditional signal processing practice. As it will be shown later, second-order descriptions do not provide adequate information to process non-Gaussian signals. One popular approach is to rely on *higher-order* statistics that exploit moments of order greater than two. If they exist, higher-order statistics provide information that is unaccessible to second-order moments [148]. Unfortunately, higher-order statistics become less reliable in impulsive environments to the extent that often they cease to exist.

The inadequacy of second- or higher-order moments leads to the introduction of alternate moment characterizations of impulsive processes. One approach is to use *fractional lower-order statistics* (FLOS) consisting of moments for orders less than two [136, 149]. Fractional lower-order statistics are not the only choice. Much like the Gaussian model naturally leads to second-order based methods, selecting a Laplacian model will lead to a different natural moment characterization. Likewise, adopting the stable laws will lead to a different, yet natural, moment characterization.

Part I

Statistical Foundations

This Page Intentionally Left Blank

2

NonGaussian Models

The Gaussian distribution model is widely accepted in signal processing practice. Theoretically justified by the Central Limit Theorem, the Gaussian model has attained a privileged place in statistics and engineering. There are, however, applications where the underlying random processes do not follow Gaussian statistics. Often, the processes encountered in practice are impulsive in nature and are not well described with conventional Gaussian distributions. Traditionally, the design emphasis has often relied on a continuity principle: optimal processing at the ideal Gaussian model should be almost optimal nearby. Unfortunately, this reliance on continuity is unfounded and in many cases one finds that optimum signal-processing methods can suffer drastic performance degradations, even for small deviations from the nominal assumptions. As an example, synchronization, detection, and equalization, basic in all communication systems, fail in impulsive noise environments whenever linear processing is used.

In order to model nonGaussian processes, a wide variety of distributions with heavier-than-Gaussian tails have been proposed as viable alternatives. This chapter reviews several of these approaches and focuses on two distribution families, namely the class of generalized Gaussian distributions and the class of stable distributions. These two distribution families are parsimonious in their characterization leading to a balanced trade-off between fidelity and complexity. On the one hand, fidelity leads to more efficient signal-processing algorithms, while the complexity issue stands for simpler models from which more tractable algorithms can be derived. The Laplacian distribution, a special case of the generalized Gaussian distribution, lays the statistical foundation for a large class of signal-processing algorithms based on the sample median. Likewise, signal processing based on the so-called sample myriad emerges from the statistical foundation laid by stable distributions.

2.1 GENERALIZED GAUSSIAN DISTRIBUTIONS

The Central Limit Theorem provides a theoretical justification for the appearance of Gaussian processes in nature. Intimately related to the Gaussian model are linear estimation methods and, to a large extent, a large section of signal-processing algorithms based on operations satisfying the linearity property. While the Central Limit Theorem has provided the key to understanding the interaction of a large number of random independent events, it has also provided the theoretical burden favoring the use of linear methods, even in circumstances where the nature of the underlying signals are decidedly non-Gaussian.

One approach used in the modeling of non-Gaussian processes is to start from the Gaussian model and slightly modify it to account for the appearance of clearly inappropriate samples or outliers. The *Gaussian mixture* or *contaminated Gaussian* model follows this approach, where the ϵ-contaminated density function takes on the form

$$f(x) = (1-\epsilon)f_n(x) + \epsilon f_c(x) \tag{2.1}$$

where $f_n(x)$ is the *nominal* Gaussian density with variance σ_n^2, ϵ is a small positive constant determining the percentage of contamination, and $f_c(x)$ is the contaminating Gaussian density with a large relative variance, such that $\sigma_c^2 \gg \sigma_n^2$. Intuitively, one out of $1/\epsilon$ samples is allowed to be contaminated by the higher variance source. The advantage of the contaminated Gaussian distribution lies in its mathematical simplicity and ease of computer simulation. Gaussian mixtures, however, present drawbacks. First, dispersion and impulsiveness are characterized by three parameters, $\epsilon, \sigma_n, \sigma_c$, which may be considered overparameterized. The second drawback, and perhaps the most serious, is that its sum density function formulation makes it difficult to manipulate in general estimation problems.

A more accurate model for impulsive phenomena was proposed by Middleton (1977) [143]. His class A, B, and C models are perhaps the most credited statistical-physical characterization of radio noise. These models have a direct physical interpretation and have been found to provide good fits to a variety of noise and interference measurements. Contaminated Gaussian mixtures can in fact be derived as approximations to Middleton's Class A model. Much like Gaussian mixtures, however, Middleton's models are complicated and somewhat difficult to use in laying the foundation of estimation algorithms.

Among the various extensions of the Gaussian distributions, the most popular models are those characterized by the *generalized Gaussian distribution*. These have been long known, with references dating back to 1923 by Subbotin [183] and 1924 by Frèchet [74]. A special case of the generalized Gaussian distribution class is the well known Laplacian distribution, which has even older roots; Laplace introduced it more than two hundred years ago [122]. In the generalized Gaussian distribution, the presence of outlier samples can be modeled by modifying the Gaussian distribution, allowing the exponential rate of tail decay to be a free parameter. In this manner, the tail of the generalized Gaussian density function is governed by the parameter k.

DEFINITION 2.1 (GENERALIZED GAUSSIAN DISTRIBUTION) *The probability density function for the generalized Gaussian distribution is given by*

$$f(x) = \frac{k\alpha}{2\Gamma(1/k)} e^{-(\alpha|x-\beta|)^k}, \tag{2.2}$$

where $\Gamma(\cdot)$ *is the Gamma function* $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$*,* α *is a constant defined as* $\alpha = \sigma^{-1}\sqrt{\Gamma(3/k)\left(\Gamma(1/k)\right)^{-1}}$ *and* σ *is the standard deviation*[1].

In this representation, the scale of the distribution is determined by the parameter $\sigma > 0$ whereas the impulsiveness is related to the parameter $k > 0$. As expected, the representation in (2.2) includes the standard Gaussian distribution as a special case for $k = 2$. Conceptually, the lower the value of k, the more impulsive the distribution is. For $k < 2$, the tails decay slower than in the Gaussian case, resulting in a heavier tailed distribution. A second special case of the generalized Gaussian distribution that is of particular interest is the case $k = 1$, which yields the double exponential, or Laplacian distribution ,

$$f(x) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{\sqrt{2}}{\sigma}|x-\beta|} = \frac{\lambda}{2} e^{-\lambda|x-\beta|}. \tag{2.3}$$

where the second representation is the most commonly used and is obtained making $\sigma = \sqrt{2}/\lambda$.

The effect of decreasing k on the tails of the distribution can be seen in Figures 2.1 and 2.2. As these figures show, the Laplacian distribution has heavier tails than the Gaussian distribution. One of the weaknesses of the generalized Gaussian distribution is the shape of these distributions around the origin for $k < 2$. The "peaky" shape of these distributions contradicts the widely accepted *Winsor's principle*, according to which, all density functions of practical appeal are bell-shaped [87, 188].

2.2 STABLE DISTRIBUTIONS

Stable distributions describe a rich class of processes that allow heavy tails and skewness. The class was characterized by Lévy in 1925 [128]. Stable distributions are described by four parameters: an *index of stability* $\alpha \in (0, 2]$, a scale parameter $\gamma > 0$, a skewness parameter $\delta \in [-1, 1]$, and a location parameter $\beta \in \mathcal{R}$. The stability

[1] The gamma function satisfies: $\Gamma(x) = (x-1)\Gamma(x-1)$ for $x > 1$. For positive integers it follows that $\Gamma(x) = (x-1)!$ and for a non integer $x > 0$ such that $x = i + u$ where $0 \le u < 1$, $\Gamma(x) = (x-1)(x-2)\cdots u\Gamma(u)$. For $x = \frac{1}{2}$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

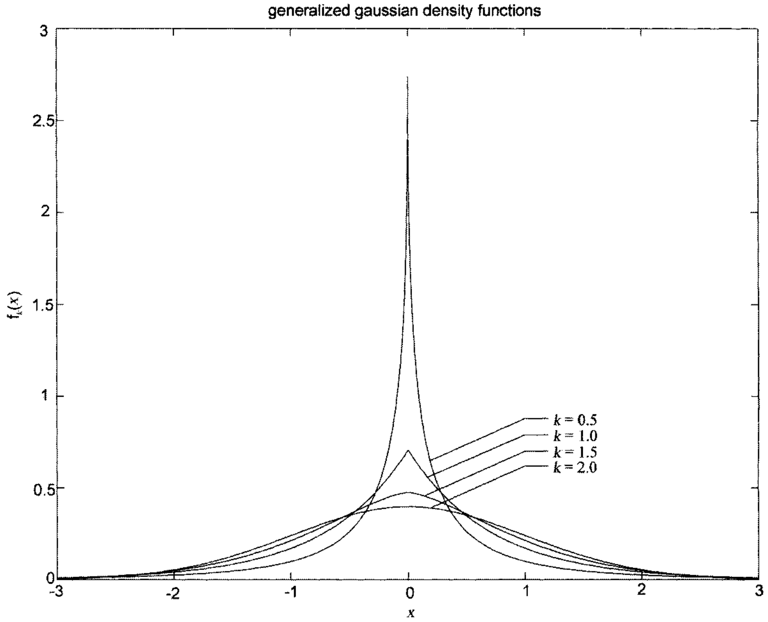


Figure 2.1 Generalized Gaussian density functions for different values of the tail constant k.

parameter α measures the thickness of the tails of the distribution and provides this model with the flexibility needed to characterize a wide range of impulsive processes. The scale parameter γ, also called the dispersion, is similar to the variance of the Gaussian distribution. The variance equals twice the square of gamma in the Gaussian case when $\alpha = 2$. When the skewness parameter is set to $\delta = 0$, the stable distribution is symmetric about the location parameter β. Symmetric stable processes are also referred to as symmetric α-stable or simply as SαS. A stable distribution with parameter α is said to be *standard* if $\beta = 0$ and $\gamma = 1$. For any stable variable X with parameters $\alpha, \beta, \gamma, \delta$, the corresponding standardized stable variable is found as $(X - \beta)/\gamma$, for $\alpha \neq 1$.

Stable distributions are rapidly becoming popular for the characterization of impulsive processes for the following reasons. Firstly, good empirical fits are often found using stable distributions on data exhibiting skewness and heavy tails. Secondly, there is solid theoretical justification that nonGaussian stable processes emerge in practice, such as multiple access interference in a Poisson-distributed communication network [179], reflection off a rotating mirror [69], and Internet traffic [127]; see Uchaikin and Zolotarev (1999) [191] and Feller (1971) [69] for additional examples. The third argument for modeling with stable distributions is perhaps the most significant and compelling. Stable distributions satisfy an important generalization

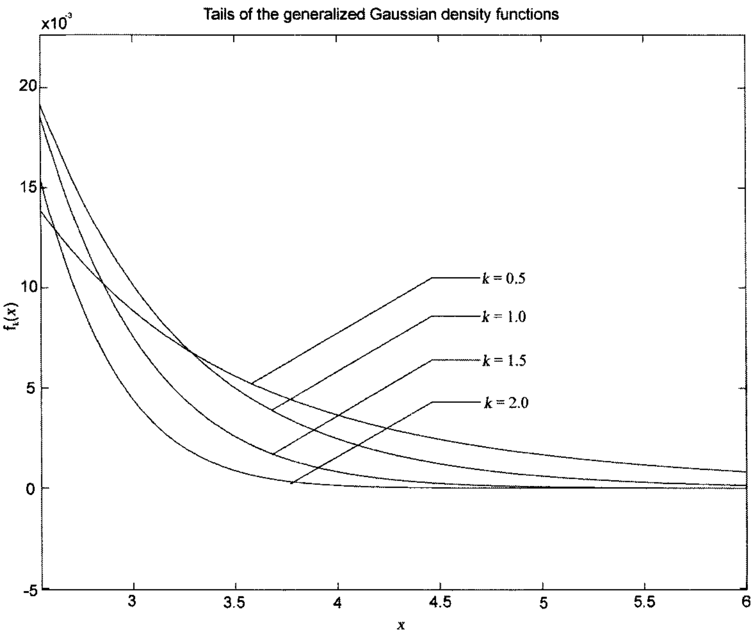


Figure 2.2 Tails of the Generalized Gaussian density functions for different values of the tail constant k.

of the Central Limit Theorem which states that the only possible limit of normalized sums of independent and identically distributed terms is stable.

A wide variety of impulsive processes found in signal processing applications arise as the superposition of many small independent effects. While Gaussian models are clearly inappropriate, stable distributions have the theoretical underpinnings to accurately model these type of impulsive processes [149, 207]. Stable models are thus appealing, since the generalization of the Central Limit Theorem explains the apparent contradictions of its "ordinary" version, which could not naturally explain the presence of heavy tailed signals.

The Generalized Central Limit Theorem and the strong empirical evidence is used by many to justify the use of stable models. Examples in finance and economics are given in Mandelbrot (1963) [138] and McCulloch (1966) [142]; in communication systems by Stuck and Kleiner (1974)[182], Nikias and Shao (1995) [149], and Ilow and Hatzinakos (1997) [106]. A number of monographs providing in-depth discussion of stable processes have recently appeared: Zolotarev (1986) [207], Samorodnitsky and Taqqu (1994) [75], Nikias and Shao (1995) [149], Uchaikin and Zolotarev (1999) [191], Adler et al. (2002) [67], and Nolan (2002) [151].

2.2.1 Definitions

Gaussian random variables obey the important property that the sum of any two Gaussian variables is itself a Gaussian random variable. Formally, for any two independent Gaussian random variables X_1 and X_2 and any positive constants a, b, c,

$$aX_1 \ + \ bX_2 \overset{d}{=} cX + d,$$

where d is a real-valued constant[2]. As their name implies, stable random variables obey this property as well.

DEFINITION 2.2 (STABLE RANDOM VARIABLES) *A random variable X is* stable *if for X_1 and X_2 independent copies of X and for arbitrary positive constants a and b, there are constants c and d such that*

$$aX_1 \ + \ bX_2 \overset{d}{=} cX + d. \tag{2.4}$$

A symmetric *stable random variable distributed around* 0 *satisfies $X \overset{d}{=} -X$.*

Informally, the stability property states that the shape of X is preserved under addition up to scale and shift. The stability property (2.4) for Gaussian random variables can be readily verified yielding $c^2 = a^2 + b^2$ and $d = (a + b - c)\mu$, where μ is the mean of the parent Gaussian distribution. Other well known distributions that satisfy the stable property are the Cauchy and Lévy distributions, and as such, both distributions are members of the stable class. The density function, for $X \sim \text{Cauchy}(\gamma, \beta)$ has the form

$$f(x) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x - \beta)^2}, \quad -\infty < x < \infty. \tag{2.5}$$

The Lévy density function, sometimes referred to as the Pearson distribution, is totally skewed concentrating on $(0, \infty)$. The density function for $X \sim \text{Lévy}(\gamma, \beta)$ has the form

$$f(x) = \sqrt{\frac{\gamma}{2\pi}} \frac{1}{(x - \beta)^{3/2}} \exp\left(-\frac{\gamma}{2(x - \beta)}\right), \quad \beta < x < \infty. \tag{2.6}$$

Figure 2.3 shows the plots of the standardized Gaussian, Cauchy, and Lévy distributions. Both Gaussian and Cauchy distributions are symmetric and bell-shaped. The main difference between these two densities is the area under their tails — the Cauchy having much larger area or heavier tails. In contrast to the Gaussian and Cauchy, the Lévy distribution is highly skewed, with even heavier tails than the Cauchy.

General stable distributions allow for varying degrees of skewness, the influence of the parameter δ in the distribution of an α-stable random variable is shown in Figure 2.4.

[2] The symbol $\overset{d}{=}$ defines equality in distribution

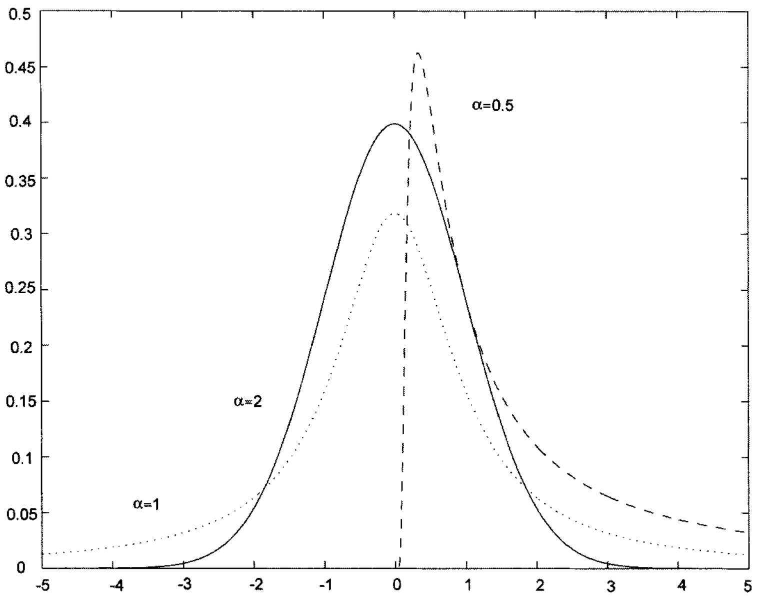


Figure 2.3 Density functions of standardized Gaussian ($\alpha = 2$), Cauchy ($\alpha = 1$), and Lévy ($\alpha = 0.5,\ \delta = 1$).

Although some practical processes might be better modeled by skewed distributions, we will focus on *symmetric* stable processes for several reasons. First, the processes found in a number of signal-processing applications are symmetric; second, asymmetric models can lead to a significant increase in the computational complexity of signal-processing algorithms; and, more important, estimating the location of an asymmetric distribution is not a well-defined problem. All of the above constitute impediments to the derivation of a general theory of nonlinear filtering.

2.2.2 Symmetric Stable Distributions

Symmetric α-stable or $S\alpha S$ distributions are defined when the skewness parameter δ is set to zero. In this case, a random variable obeying the symmetric stable distribution with scale γ is denoted as $X \sim S\alpha S(\gamma)$. Although the stability condition in Definition 2.2 is sufficient to characterize all stable distributions, a second and more practical characterization of stable random variables is through their characteristic function.

$$\phi(\omega) = E\exp(j\omega X) = \int_{-\infty}^{\infty} \exp(j\omega x) f(x)dx \tag{2.7}$$

where $f(x)$ is the density function of the underlying random variable.

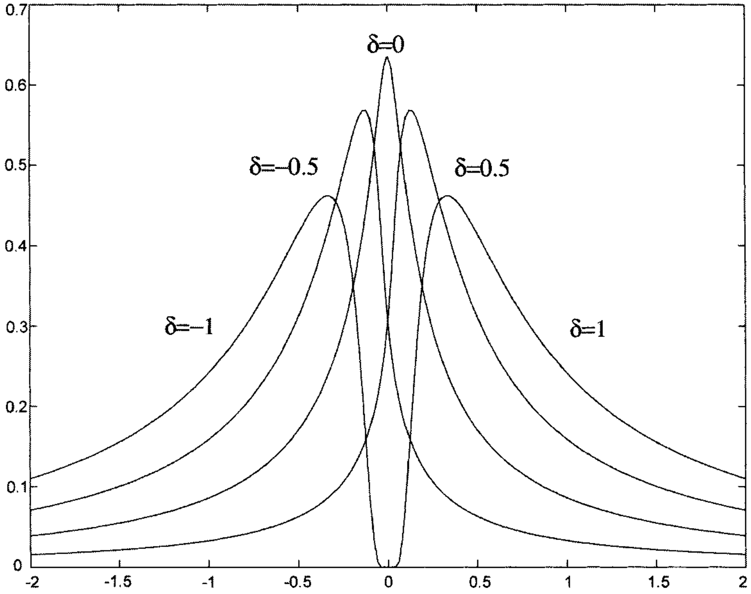


Figure 2.4 Density functions of skewed stable variables ($\alpha = 0.5,\ \gamma = 1,\ \beta = 0$).

DEFINITION 2.3 (CHARACTERISTIC FUNCTION OF SαS DISTRIBUTIONS) *A random variable X is* symmetrically stable *if and only if $X \overset{d}{=} AZ + B$ where $0 < \alpha \leq 2, A \geq 0, B \in \mathcal{R}$ and $Z = Z(\alpha)$ is a random variable with characteristic function*

$$\phi(\omega) = e^{-\gamma^{\alpha}|\omega|^{\alpha}}. \tag{2.8}$$

The dispersion parameter γ is a positive constant related to the scale of the distribution. Again, the parameter α is referred to as the index of stability. In order for (2.8) to define a characteristic function, the values of α must be restricted to the interval $(0; 2]$. Conceptually speaking, α determines the impulsiveness or tail heaviness of the distribution (smaller values of α indicate increased levels of impulsiveness). The limit case, $\alpha = 2$, corresponds to the zero-mean Gaussian distribution with variance $2\gamma^2$.[3] All other values of α correspond to heavy-tailed distributions.

Figure 2.5 shows plots of normalized unitary-dispersion stable densities. Note that lower values of α correspond to densities with heavier tails, as shown in Figure 2.6.

[3] The characteristic function of a Gaussian random variable with zero mean and variance σ^2 is given by: $\phi(\omega) = \exp\left(-\frac{\omega^2\sigma^2}{2}\right)$, from this equation and (2.8) with $\alpha = 2$, the relationship shown between γ and σ^2 can be obtained.

Symmetric stable densities maintain many of the features of the Gaussian density. They are smooth, unimodal, symmetric with respect to the mode, and bell-shaped.

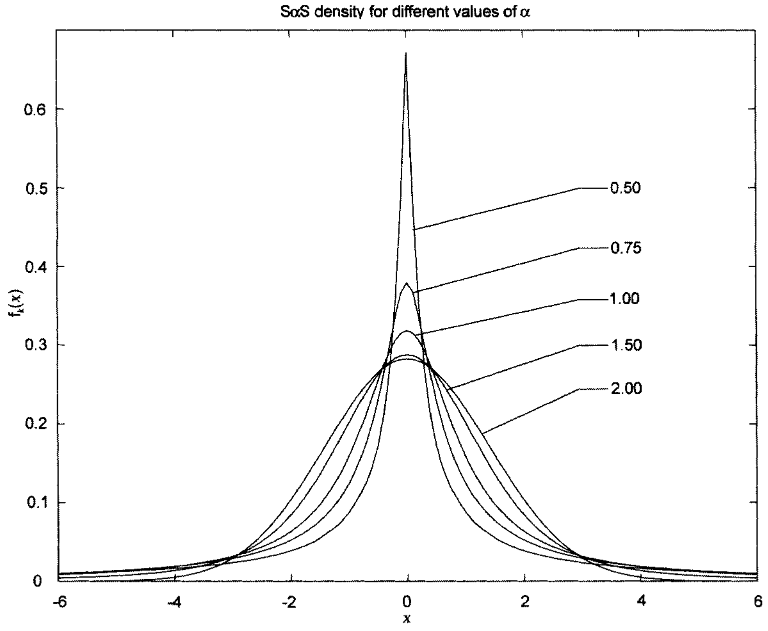


Figure 2.5 Density functions of Symmetric stable distributions for different values of the tail constant α.

A major drawback to stable distribution modeling is that with a few exceptions stable density or their corresponding cumulative distribution functions lack closed form expressions. There are three cases for which closed form expressions of stable density functions exist: the Gaussian distribution ($\alpha = 2$), the Cauchy distribution ($\alpha = 1$), and the Lévy ($\alpha = \frac{1}{2}$) distribution. For other values of α, no closed form expressions are known for the density functions, making it necessary to resort to series expansions or integral transforms to describe them.

DEFINITION 2.4 (SYMMETRIC STABLE DENSITY FUNCTIONS) *A general, "zero-centered," symmetric stable random variable with unitary dispersion can be characterized by the power series density function representation [207]:*

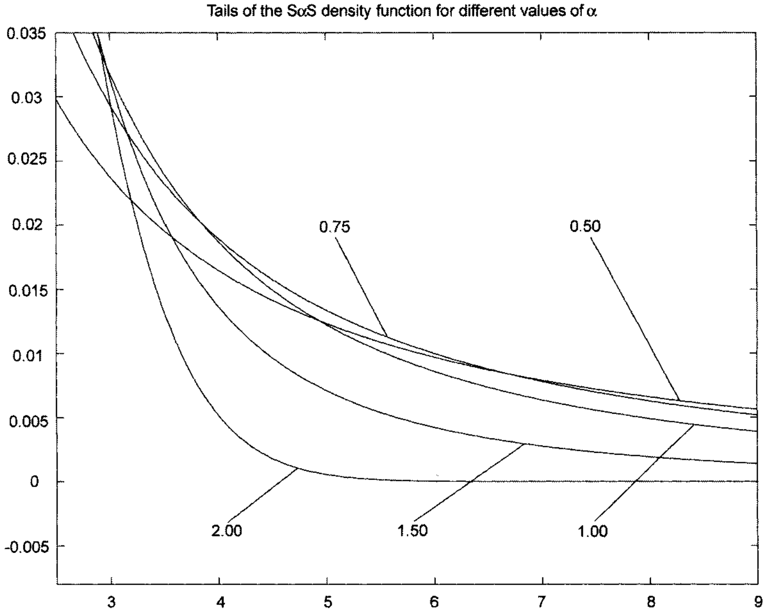


Figure 2.6 Tails of symmetric stable distributions for different values of the tail constant α.

$$f_\alpha(x) = \begin{cases} \frac{1}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{k!}\Gamma(k\alpha+1)\sin(\frac{\pi k\alpha}{2})|x|^{-k\alpha-1} & \textit{for}\ 0<\alpha<1,\ x\neq 0 \\ \frac{1}{\pi(x^2+1)} & \textit{for}\ \alpha=1 \\ \frac{1}{\pi\alpha}\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k)!}\Gamma(\frac{2k+1}{\alpha})x^{2k} & \textit{for}\ 1<\alpha<2 \\ \frac{1}{2\sqrt{\pi}}\exp[-\frac{x^2}{4}] & \textit{for}\ \alpha=2. \end{cases} \tag{2.9}$$

DEFINITION 2.5 (CHARACTERISTIC FUNCTION OF A STABLE RV.) *[151] A random variable X is stable with characteristic exponent α, dispersion γ, location β and skewness δ if X has a characteristic function:*

$$\phi(\omega) = \begin{cases} \exp\left(-\gamma^\alpha|\omega|^\alpha\left[1-j\delta\left(\tan\frac{\pi\alpha}{2}\right)(sgn\ \omega)\right]+j\beta\omega\right) & \textit{for}\alpha\neq 1 \\ \exp\left(-\gamma|\omega|\left[1+j\delta\frac{2}{\pi}(sgn\ \omega)\ln|\omega|\right]+j\beta\omega\right) & \textit{for}\alpha=1. \end{cases} \tag{2.10}$$

EXAMPLE 2.1 (STANDARD STABLE RANDOM VARIABLES)

As stated previously, if X is a stable random variable with location β and dispersion γ, the variable $X' = \frac{X-\beta}{\gamma}$ $(\alpha \neq 1)$ is standard stable. This can be demonstrated with the help of the general characteristic function. Define

$$\begin{aligned}
\phi(\omega') &= E\left[\exp\left(j\omega' X'\right)\right] = E\left[\exp\left(j\omega'\frac{X-\beta}{\gamma}\right)\right] \\
&= \exp\left(-j\frac{\omega'}{\gamma}\beta\right) E\left[\exp\left(j\frac{\omega'}{\gamma}X\right)\right] \quad \text{using (2.10)} \\
&= \exp\left(-j\frac{\omega'}{\gamma}\beta\right)\exp\left(-\gamma^\alpha\left|\frac{\omega'}{\gamma}\right|^\alpha\left[1-j\delta\left(\tan\frac{\pi\alpha}{2}\right)\left(\operatorname{sgn}\frac{\omega'}{\gamma}\right)\right]+j\beta\frac{\omega'}{\gamma}\right)
\end{aligned}$$

but $\gamma \geq 0$, then $|\gamma| = \gamma$ and $\operatorname{sgn}\left(\frac{\omega'}{\gamma}\right) = \operatorname{sgn}(\omega')$, then

$$\phi(\omega') = \exp\left(-|\omega|^\alpha\left[1-j\delta\left(\tan\frac{\pi\alpha}{2}\right)(\operatorname{sgn}\omega')\right]\right) \tag{2.11}$$

is the characteristic function of a stable random variable with $\gamma = 1$ and $\beta = 0$. ■

EXAMPLE 2.2

Let $X \sim S(\alpha, \gamma, \beta)$, a symmetric stable random variable, then for $a \neq 0$ it is shown that

$$aX + b \sim S(\alpha, |a|\gamma, a\beta + b).$$

Following the procedure used in the previous example, define $X' = aX + b$:

$$\begin{aligned}
\phi(\omega') &= E\left[\exp\left(j\omega' X'\right)\right] = E\left[\exp\left(j\omega'\left(aX+b\right)\right)\right] \\
&= \exp\left(j\omega' b\right) E\left[\exp\left(j\left(\omega' a\right)X\right)\right] \quad \text{using (2.10) with } \delta = 0 \\
&= \exp\left(j\omega' b\right)\exp\left(-\gamma^\alpha\left|\omega' a\right|^\alpha + j\beta\left(\omega' a\right)\right) \\
&= \exp\left(-\left(|a|\gamma\right)^\alpha\left|\omega'\right|^\alpha + j\omega'\left(a\beta+b\right)\right),
\end{aligned} \tag{2.12}$$

which is the characteristic function of a symmetric stable random variable with dispersion $|a|\gamma$ and location $a\beta + b$. ■

EXAMPLE 2.3

Let $X_1 \sim S(\alpha, \gamma_1, \beta_1)$ and $X_2 \sim S(\alpha, \gamma_2, \beta_2)$ be independent symmetric stable random variables, it is shown here that $X_1 + X_2 \sim S(\alpha, \gamma, \beta)$, where $\gamma^\alpha = \gamma_1^\alpha + \gamma_2^\alpha$ and $\beta = \beta_1 + \beta_2$.

Define $X' = X_1 + X_2$ and find the characteristic function of X' as:

$$\begin{aligned}
\phi(\omega') &= E\left[\exp\left(j\omega' X'\right)\right] = E\left[\exp\left(j\omega'\left(X_1 + X_2\right)\right)\right] \\
&= E\left[\exp\left(j\omega' X_1\right)\right] E\left[\exp\left(j\omega' X_2\right)\right] \quad \text{since the variables are independent} \\
&= \exp\left(-\gamma_1^\alpha \left|\omega'\right|^\alpha + j\beta_1\omega'\right) \exp\left(-\gamma_2^\alpha \left|\omega'\right|^\alpha + j\beta_2\omega'\right) \\
&= \exp\left(-\left(\gamma_1^\alpha + \gamma_2^\alpha\right)\left|\omega'\right|^\alpha + j(\beta_1 + \beta_2)\omega'\right), \qquad (2.13)
\end{aligned}$$

which is the characteristic function of a symmetric stable random variable with $\gamma^\alpha = \gamma_1^\alpha + \gamma_2^\alpha$ and $\beta = \beta_1 + \beta_2$. ■

2.2.3 Generalized Central Limit Theorem

Much like Gaussian signals, a wide variety of non-Gaussian processes found in practice arise as the superposition of many small independent effects. At first, this may point to a contradiction of the Central Limit Theorem, which states that, in the limit, the sum of such effects tends to a Gaussian process. A careful revision of the conditions of the Central Limit Theorem indicates that, in order for the Central Limit Theorem to be valid, the variance of the superimposed random variables must be finite. If the variance of the underlying random variables is infinite, an important generalization of the Central Limit Theorem emerges. This generalization explains the apparent contradictions of its "ordinary" version, as well as the presence of non-Gaussian processes in practice.

THEOREM 2.1 (GENERALIZED CENTRAL LIMIT THEOREM [75]) *Let X_1, $X_2, \ldots$ be an independent, identically distributed sequence of (possibly shift corrected) random variables. There exist constants a_n such that as $n \rightarrow \infty$ the sum*

$$a_n(X_1 + X_2 + \cdots) \xrightarrow{d} Z \qquad (2.14)$$

if and only if Z is a stable random variable with some $0 < \alpha \leq 2$.

In the same way as the Gaussian model owes most of its strength to the Central Limit Theorem, the Generalized Central Limit Theorem constitutes a strong theoretical argument compelling the use of stable models in practical problems.

At first, the use of infinite variance in the definition of the Generalized Central Limit Theorem may lead to some skepticism as infinite variance for real data having bounded range may seem inappropriate. It should be noted, however, that the variance is but one measure of spread of a distribution, and is not appropriate for all problems. It is argued that in stable environments, γ may be more appropriate as a measure of spread. From an applied point of view, what is important is capturing the shape of a distribution. The Gaussian distribution is, for instance, routinely used to model bounded data, even though it has unbounded support. Although in some cases there are solid theoretical reasons for believing that a stable model is appropriate, in other more pragmatic cases the stable model can be used if it provides a good and parsimonious fit to the data at hand.

2.2.4 Simulation of Stable Sequences

Computer simulation of random processes is important in the design and analysis of signal processing algorithms. To this end, Chambers, Mallows, and Stuck (1976) [43] developed an algorithm for the generation of stable random variables. The algorithm is described in the following theorem.

THEOREM 2.2 (SIMULATION OF STABLE VARIABLES [151]) *Let Θ and W be independent with Θ uniformly distributed on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and W exponentially distributed with mean* 1. *$Z \sim S(\alpha, \delta)$ is generated as*

$$Z = \begin{cases} c(\alpha,\delta) \frac{\sin \alpha(\Theta+\theta_0)}{(\cos\Theta)^{1/\alpha}} \left(\frac{cos(\Theta-\alpha(\Theta+\theta_0))}{W}\right)^{(1-\alpha)/\alpha} & \alpha \neq 1 \\ \frac{2}{\pi}\left[\left(\frac{\pi}{2}+\delta\Theta\right)\tan\Theta - \delta\ln\left(\frac{\frac{\pi}{2}W\cos\Theta}{\frac{\pi}{2}+\delta\Theta}\right)\right] & \alpha = 1 \end{cases} \qquad (2.15)$$

where $c(\alpha,\delta) = (1 + (\delta \tan \frac{\pi\alpha}{2})^2)^{1/(2\alpha)}$ and $\theta_0 = \alpha^{-1} \arctan(\delta \tan \frac{\pi\alpha}{2})$. In particular, for $\alpha = 1, \delta = 0$ (Cauchy), $Z \sim \mathrm{Cauchy}(\gamma)$ is generated as

$$Z = \gamma \tan(\Theta) = \gamma \tan\left(\pi\left(U - \frac{1}{2}\right)\right) \qquad (2.16)$$

where U is a uniform random variable in $(0, 1)$.

Figure 2.7 illustrates the impulsive behavior of symmetric stable processes as the characteristic exponent α is varied. Each one of the plots shows an independent and identically distributed (i.i.d.) “zero-centered” symmetric stable signal with unitary geometric power[4]. In order to give a better feeling of the impulsive structure of the data, the signals are plotted twice under two different scales. As it can be appreciated, the Gaussian signal ($\alpha = 2$) does not show impulsive behavior. For values of α close to 2 ($\alpha = 1.7$ in the figure), the structure of the signal is still similar to the Gaussian,

[4] The geometric power is introduced in the next section as a strength indicator of processes with infinite variance.

although some impulsiveness can now be observed. As the value of α is decreased, the impulsive behavior increases progressively.

2.3 LOWER-ORDER MOMENTS

Statistical signal processing relies, to a large extent, on the statistical characterization provided by second-order moments such as the variance $Var(X) = E(X^2) - (EX)^2$ with EX being the first moment. Second-order based estimation methods are sufficient whenever the underlying signals obey Gaussian statistics. The characterization of nonGaussian processes by second-order moments is no longer optimal and other moment characterizations may be required. To this end, ***higher-order statistics*** (HOS) exploiting third- and fourth-order moments (cummulants) have led to improved estimation algorithms in nonGaussian environments, provided that higher-order moments exist and are finite [148]. In applications where the processes are inherently impulsive, second-order and HOS may either be unreliable or may not even exist.

2.3.1 Fractional Lower-Order Moments

The different behavior of the Gaussian and nonGaussian distributions is to a large extent caused by the characteristics of their tails. The existence of second-order moments depends on the behavior of the tail of the distribution. The tail "thickness" of a distribution can be measured by its asymptotic mass $P(|X| > x)$ as $x \to \infty$. Given two functions $h(x)$ and $g(x)$, they have asymptotic similarity ($h(x) \sim g(x)$) if for $x \to \infty$: $\lim_{x\to\infty} h(x)/g(x) = 1$, the Gaussian distribution can be shown to have exponential order tails with asymptotic similarity

$$P(|X| > x) \sim \sqrt{\frac{2}{\pi}} x^{-1} e^{-x^2/2}. \tag{2.17}$$

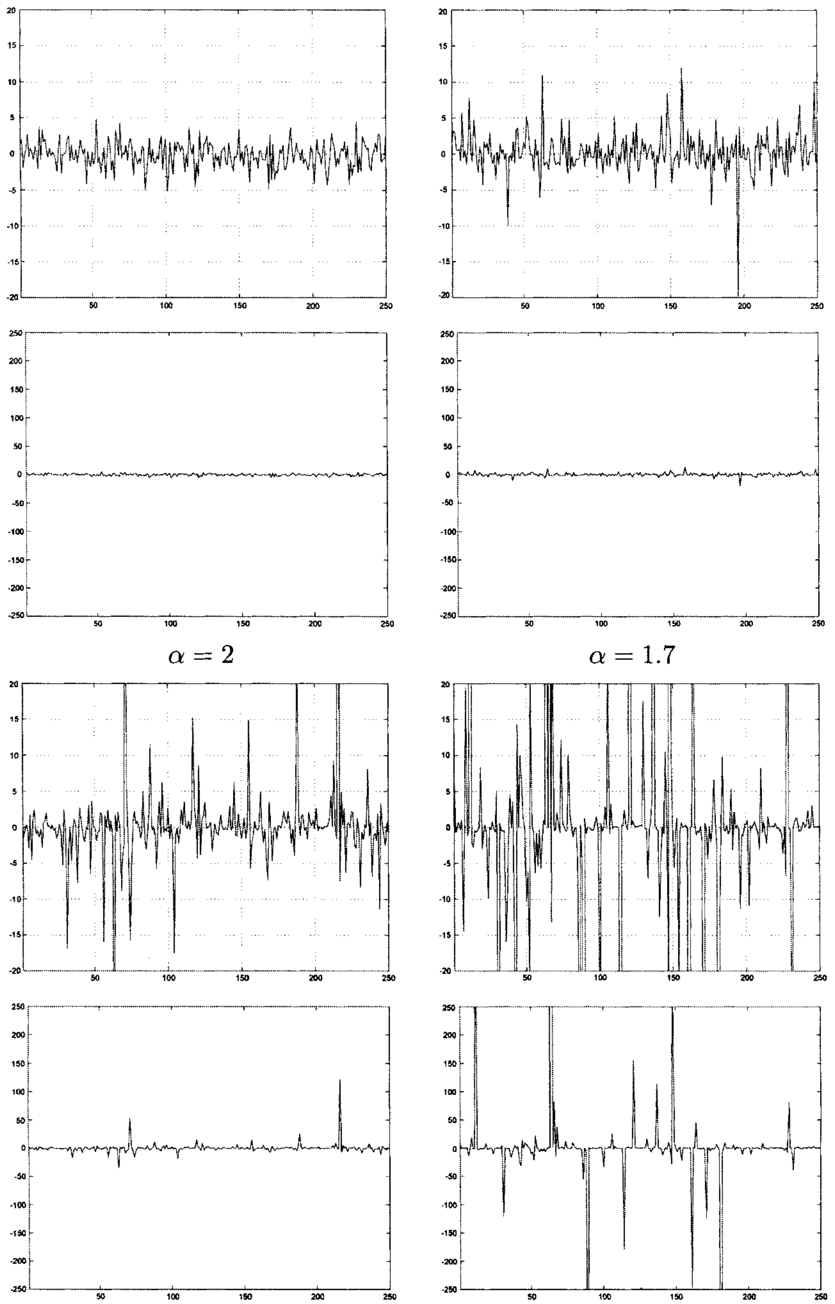
Second order moments for the Gaussian distribution are thus well behaved due to the exponential order of the tails. The tails of the Laplacian distribution are heavier than that of the Gaussian distribution but remain of exponential order with

$$P(|X| > x) \sim e^{-x/\sigma}. \tag{2.18}$$

The tails of more impulsive nonGaussian distributions, however, behave very differently. Infinite variance processes that can appear in practice as a consequence of the Generalized Central Limit Theorem are modeled by probability distributions with algebraic tails for which

$$P(X > x) \sim cx^{-\alpha} \tag{2.19}$$

for some fixed c and $\alpha > 0$. The tail-heaviness of these distributions is determined by the tail constant α, with increased impulsiveness corresponding to small values of α. Stable random variables, for $\alpha < 2$, are examples of processes having algebraic tails as described by the following theorem.



$\alpha = 2$

$\alpha = 1.7$

$\alpha = 1$

$\alpha = 0.6$

Figure 2.7 Impulsive behavior of i.i.d. α-stable signals as the tail constant α is varied. Signals are plotted twice under two different scales.

THEOREM 2.3 (STABLE DISTRIBUTION TAILS [151]) *Let $X \sim S(\alpha)$ be a symmetric, standard stable random variable with* $0 < \alpha < 2$, *then as* $x \rightarrow \infty$,

$$P(X > x) \sim \Gamma(\alpha)\frac{sin(\pi\alpha/2)}{\pi}x^{-\alpha}. \tag{2.20}$$

For stable and other distributions having algebraic tails, the following theorem is important having a significant impact on the statistical moments that can be used to process and analyze these signals.

THEOREM 2.4 *Algebraic-tailed random variables exhibit finite absolute moments for orders less than* α

$$E(|X|^p) < \infty, \quad \text{if } p < \alpha. \tag{2.21}$$

Conversely, if $p \geq \alpha$, *the absolute moments become infinite.*

Proof: The variable Y is replaced by $|X|^p$ in the first moment relationship

$$EY = \int_{-\infty}^{\infty} P(Y > y)dy \tag{2.22}$$

yielding

$$\begin{aligned} E(|X|^p) &= \int_0^{\infty} P(|X|^p > t)dt && (2.23)\\ &= \int_0^{\infty} pu^{p-1}P(|X| > u)du, && (2.24) \end{aligned}$$

which, from (2.19), diverges for any distribution having algebraic tails. ■

Given that second-order, or higher-order moments, do not exist for algebraic tailed processes, the result in (2.21) points to the fact that in this case, it is better to rely on *fractional lower-order moments* (FLOMs): $E|X|^p = \int_{-\infty}^{\infty} |x|^p f(x)dx$, which exist for $0 < p < \alpha$. FLOMs for arbitrary processes can be computed from the definitions. Zolotarev (1957) [207], for instance, derived the FLOMs of $S\alpha S$ random variables as

PROPERTY 2.1 *The FLOMs for a* $S\alpha S$ *random variable with zero location parameter and dispersion* γ *is given by*

$$E(|X|^p) = C(p,\alpha)\gamma^p \quad 0 < p < \infty, \tag{2.25}$$

where

$$C(p,\alpha) = \frac{2^{p+1}\Gamma\left(\frac{p+1}{2}\right)\Gamma(-p/\alpha)}{\alpha\sqrt{\pi}\Gamma(-p/2)}. \tag{2.26}$$

Figure 2.8 depict the fractional lower-order moments for standardized $S\alpha S$ ($\gamma = 1,\ \delta = 0$) as functions of p for various values of α.

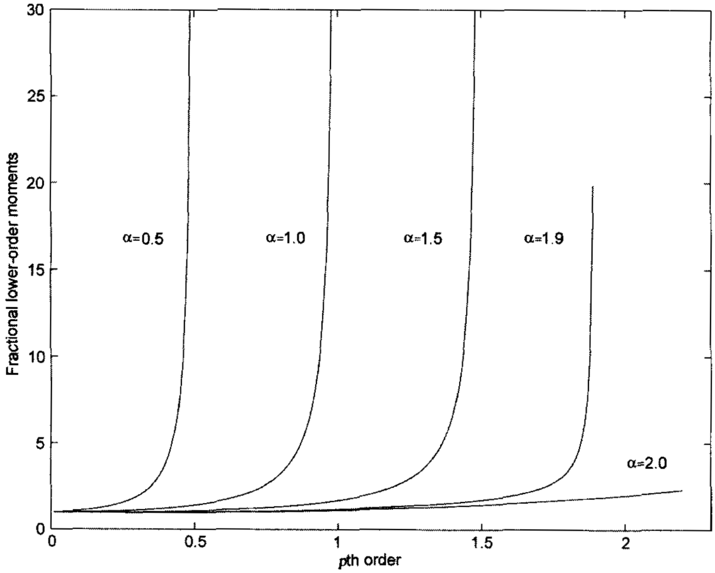


Figure 2.8 Fractional lower-order moments of the standardized $S\alpha S$ random variable.

2.3.2 Zero-Order Statistics

Fractional lower-order statistics do not provide a universal framework for the characterization of algebraic-tailed processes: for a given $p > 0$, there will always be a "remaining" class of processes (those with $\alpha \leq p$) for which the associated FLOMs do not exist. On the other hand, restricting the values of p to the valid interval $(0; \alpha)$ requires either the previous knowledge of α or a numerical procedure to estimate it. The former may not be possible in most practical applications, and the later may be inexact and/or computationally expensive. Unlike lower- or higher-order statistics, the advantage of *zero-order statistics* (ZOS) is that they provide a common ground for the analysis of basically *any* distribution of practical use [85, 48, 47, 50, 49]. In the same way as *pth*-order moments constitute the basis of FLOS and HOS techniques, zero-order statistics are based on logarithmic "moments" of the form $E \log |X|$.

THEOREM 2.5 *Let X be a random variable with algebraic or lighter tails. Then, $E \log |X| < \infty$.*

Proof: If X has algebraic or lighter tails, there exists a $p > 0$ such that $E|X|^p < \infty$. Jensen's inequality [65] guarantees that for a concave function ϕ, and a random variable Z, $E\phi(Z) \leq \phi(EZ)$. Letting $\phi(x) = \log |x|/p$ and $Z = |X|^p$ leads to

$$E \log |X| = E\left(\frac{\log |X|^p}{p}\right) \leq \frac{\log(E|X|^p)}{p} < \infty, \tag{2.27}$$

which is the desired result. ■

Random processes for which Theorem 2.5 applies, are referred to as being of "logarithmic order," in analogy with the term "second order" used to denote processes with finite variance. The logarithmic moment, which is finite for all logarithmic-order processes, can be used as a tool to characterize these signals. The strength of a signal is one attribute that can be characterized by logarithmic moments. For second-order processes, the *power* EX^2 is a widely accepted measure of signal strength. This measure, however, is always infinite when the processes exhibit algebraic tails, failing to provide useful information. To this end, zero-order statistics can be used to define an alternative strength measure referred to as the *geometric power.*

DEFINITION 2.6 (GEOMETRIC POWER [85]) *Let X be a logarithmic-order random variable. The* geometric power *of X is defined as*

$$S_0 = S_0(X) = e^{E\log|X|}. \tag{2.28}$$

The geometric power gives a useful strength characterization along the class of logarithmic-order processes having the advantage that it is mathematically and conceptually simple. In addition, it has a rich set of properties that can be effectively used. The geometric power is a scale parameter satisfying $S_0(X) \geq 0$ and $S_0(cX) = |c|S_0(X)$, and as such, it can be effectively used as an indicator of process strength or "power" in situations where second-order methods are inadequate. The geometric power takes on the value $S_0(X) = 0$ if and only if $P(X = 0) > 0$, which implies that zero power is only attained when there is a discrete probability mass located in zero [85].

The geometric power of any logarithmic-order process can be computed by the evaluation of (2.28). The geometric power of symmetric stable random variables, for instance, can be obtained in the closed-form expression.

PROPOSITION 2.1 (GEOMETRIC POWER OF STABLE PROCESSES) *The geometric power of a symmetric stable variable is given by*

$$S_0 = \frac{\gamma C_g^{1/\alpha}}{C_g}, \tag{2.29}$$

where $C_g = e^{C_e} \approx 1.78$, is the exponential of the Euler constant.

Proof: From [207], p. 215, the logarithmic moment of a zero-centered symmetric α-stable random variable with unitary dispersion is given by

$$E\log|X| = \left(\frac{1}{\alpha} - 1\right)C_e, \tag{2.30}$$

where $C_e = 0.5772\ldots$ is the Euler constant. This gives

$$S_0(X)\Bigg|_{\gamma=1} = e^{E\log|X|} = \left(e^{C_e}\right)^{\frac{1}{\alpha}-1} = \frac{C_g^{1/\alpha}}{C_g}, \tag{2.31}$$

where $C_g = e^{C_e} \approx 1.78$. If X has a non-unitary dispersion γ, it is easy to see that

$$S_0(X) = \gamma \left[S_0(X)|_{\gamma=1} \right] = \frac{\gamma C_g^{1/\alpha}}{C_g}. \tag{2.32}$$

■

The geometric power is well defined in the class of stable distributions for any value of $\alpha > 0$. Being a scale parameter, it is always multiple of γ and, more interestingly, it is a decreasing function of α. This is an intuitively pleasant property, since we should expect to observe more process strength when the levels of impulsiveness are increased.

Figure 2.9 illustrates the usefulness of the geometric power as an indicator of process strength in the α-stable framework. The scatter plot on the left side was generated from a stable distribution with $\alpha = 1.99$ and geometric power $S_0 = 1$. On the right-hand side, the scatter plot comes from a Gaussian distribution ($\alpha = 2$) also with unitary geometric power. After an intuitive inspection of Figure 2.9, it is reasonable to conclude that both of the generating processes possess the same strength, in accordance with the values of the geometric power. Contrarily, the values of the second-order power lead to the misleading conclusion that the process on the left is much stronger than the one on the right.

A similar example to the above can be constructed to depict the disadvantages of FLOS-based indicators of strength in the class of logarithmic-order processes. Fractional moments of order p present the same type of discontinuities as the one illustrated in Figure 2.9 for processes with tail constants close to $\alpha = p$. The geometric power, on the other side, is consistently continuous along all the range of values of α. This "universality" of the geometric power provides a general framework for comparing the strengths of any pair of logarithmic-order signals, in the same way as the (second-order) power is used in the classical framework.

The term zero-order statistics used to describe statistical measures using logarithmic moments is coined after the following relationship of the geometric power with fractional order statistics.

THEOREM 2.6 *Let $S_p = (E|X|^p)^{1/p}$ denote the scale parameter derived from the pth-order moment of X. If S_p exists for sufficiently small values of p, then*

$$S_0 = \lim_{p \to 0} S_p. \tag{2.33}$$

Furthermore, $S_0 \leq S_p$, for any $p > 0$.

Proof: It is enough to prove that $\lim_{p\to 0} \frac{1}{p} \log E|X|^p = E \log |X|$. Applying L'Hospital rule,

$$\lim_{p\to 0} \frac{\log E|X|^p}{p} = \lim_{p\to 0} \frac{d}{dp} \log E|X|^p \tag{2.34}$$

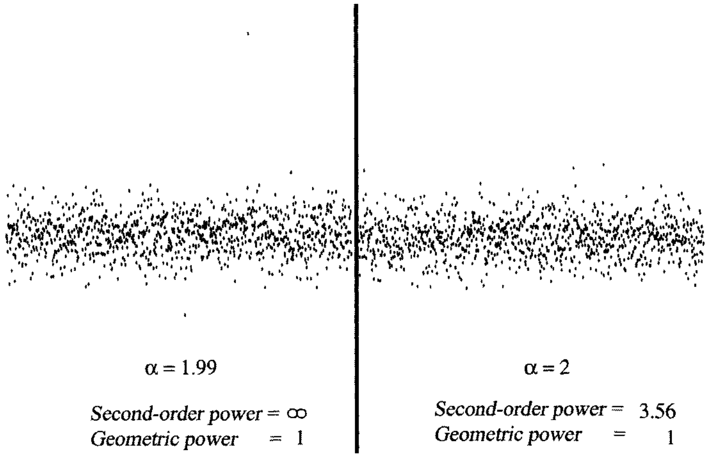


Figure 2.9 Comparison of second-order power vs. geometric power for i.i.d. α-stable processes. Left: $\alpha = 1.99$. Right: $\alpha = 2$. While the values of the geometric power give an intuitive idea of the relative strengths of the signals, second-order power can be misleading.

$$= \lim_{p\to 0} \frac{E\left(\frac{d}{dp}|X|^p\right)}{E|X|^p} \tag{2.35}$$

$$= \lim_{p\to 0} \frac{E(|X|^p \log|X|)}{E|X|^p} \tag{2.36}$$

$$= E\log|X|. \tag{2.37}$$

To prove that $S_0 \leq S_p$, Jensen's inequality [65] guarantees that for a convex function ϕ and a random variable Z, $\phi(EZ) \leq E\phi(Z)$. Making $\phi(x) = e^x$ and $Z = \log|X|^p$ we get,

$$S_0^p = e^{(E\log|X|^p)} \leq Ee^{\log|X|^p} = E|X|^p = S_p^p, \tag{2.38}$$

which leads to the desired result. ■

Theorem 2.6 indicates that techniques derived from the geometric power are the limiting *zero-order* relatives of FLOMs.

2.3.3 Parameter Estimation of Stable Distributions

The generalized central limit method and the theoretical formulation of several stochastic processes justify the use of stable distribution models. In some other cases, the approach can be more empirical where large data sets exhibit skewness and heavy tails in such fashion that stable models provide parsimonious and effective characterization. Modeling a sample set by a stable probability density function thus requires estimating the parameters of the stable distribution, namely the characteristic exponent $\alpha \in (0,2]$; the symmetry parameter $\delta \in [-1,1]$, which sets the skewness; the scale parameter $\gamma > 0$; and the location parameter β.

The often-preferred maximum likelihood parameter estimation approach, which offers asymptotic efficiency, is not readily available as stable distributions lack closed form analytical expressions. This problem can be overcome by numerical solutions. Nonetheless, simpler methods may be adequate in many cases [40, 68, 135, 151]. The approach introduced by Kuruo ğlu, in particular, is simple and provides adequate estimates in general [121]. In Kuruoğlu's approach, the data of a general α-stable distributions is first transformed to data satisfying certain symmetric and skewness conditions. The parameters of the transformed data can then be estimated by the use of simple methods that use fractional lower-order statistics. Finally, the parameter estimates of the original data are obtained by using well-known relationships between these two sets of parameters. Kuruoğlu's approach is summarized next.

Let X_k be independent α-stable variates that are identically distributed with parameters α, δ, γ, and β. This stable law is denoted as

$$X_k \sim S_\alpha(\delta, \gamma, \beta). \tag{2.39}$$

The distribution of a weighted sum of these variables with weights a_k can be derived as [121]

$$Z = \sum_{k=1}^{n} a_k X_k \sim S_\alpha\left(\frac{\sum_{k=1}^{n} a_k^{<\alpha>}}{\sum_{k=1}^{n} |a_k|^\alpha}\delta, \left(\sum_{k=1}^{n} |a_k|^\alpha\right)^{\frac{1}{\alpha}} \gamma, \sum_{k=1}^{n} a_k \beta \right) \tag{2.40}$$

where $x^{<p>}$ denotes the signed pth power of a number x

$$x^{<p>} = \text{sign}(x)|x|^p. \tag{2.41}$$

This provides a convenient way to generate sequences of independent variables with zero β, zero δ, or with zero values for both β and δ (except when $\alpha = 1$). These are referred to as the centered, deskewed, and symmetrized sequences, respectively:

$$X_k^C = X_{3k} + X_{3k-1} - 2X_{3k-2} \sim S_\alpha\left(\left[\frac{2-2^\alpha}{2+2^\alpha}\right] \delta, [2+2^\alpha]^{\frac{1}{\alpha}}\gamma, 0 \right) \tag{2.42}$$

$$X_k^D = X_{3k} + X_{3k-1} - 2^{1/\alpha} X_{3k-2} \sim S_\alpha(0, 4^{\frac{1}{\alpha}}\gamma, [2 - 2^{1/\alpha}]\beta) \tag{2.43}$$

$$X_k^S = X_{2k} - X_{2k-1} \sim S_\alpha(0, 2^{\frac{1}{\alpha}}\gamma, 0). \tag{2.44}$$

Using such simpler sequences, moment methods for parameter estimation can be easily applied for variates with $\beta = 0$ or $\delta = 0$, or both, to the general variates at the cost of loss of some sample size. In turn, these estimates are used to calculate the estimates of the original α-stable distributions.

Moments of a distribution provide important statistical information about the distribution. Kuruoğlu's methods, in particular, exploit fractional lower-order or negative-order moments, which, for the skewed α-stable distributions, are finite for certain parameter values. First, the absolute and signed fractional-order moments of stable variates are calculated analytically as a generalization of Property 2.1 [121].

PROPERTY 2.2 *Let $X \sim S_\alpha(\delta, \gamma, 0)$. Then, for $\alpha \neq 1$*

$$\mathrm{E}[|X|^p] = \frac{\Gamma\left(1-\frac{p}{\alpha}\right)}{\Gamma(1-p)} \left(\frac{\gamma}{|\cos\theta|^{\frac{1}{\alpha}}}\right)^{p} \frac{\cos\left(\frac{p\theta}{\alpha}\right)}{\cos\left(\frac{p\pi}{2}\right)} \tag{2.45}$$

for $p \in (-1, \alpha)$ and where

$$\theta = \arctan\left(\delta \tan\frac{\alpha\pi}{2}\right). \tag{2.46}$$

As for the signed fractional moment of skewed α-stable distributions, the following holds [121].

PROPERTY 2.3 *Let $X \sim S_\alpha(\delta, \gamma, 0)$. Then*

$$\mathrm{E}[X^{<p>}] = \frac{\Gamma(1-\frac{p}{\alpha})}{\Gamma(1-p)} \left(\frac{\gamma}{|\cos\theta|^{\frac{1}{\alpha}}}\right)^{p} \frac{\sin(\frac{p\theta}{\alpha})}{\sin(\frac{p\pi}{2}).} \tag{2.47}$$

Given n independent observations of a random variate X, the absolute and signed fractional moments can be estimated by the sample statistics:

$$A_p = \tfrac{1}{n}\textstyle\sum_{k=1}^{n} |X_k|^p, \quad S_p = \tfrac{1}{n}\textstyle\sum_{k=1}^{n} X_k^{<p>}. \tag{2.48}$$

The presence of the gamma function in the formulae presented by the propositions hampers the direct solution of these expressions. However, by taking products and ratios of FLOMs and applying the following property of the gamma function:

$$\Gamma(p)\Gamma(1-p) = \frac{\pi}{\sin(p\pi)} \tag{2.49}$$

a number of simple closed-form estimators for α, δ, and γ can be obtained.

FLOM estimate for α: Noting that (2.48) is only the approximation of the absolute and signed fractional order moments, the analytic formulas (2.45), (2.47) are used. From (2.45), the product $A_p A_{-p}$ is given by

$$A_p A_{-p} = \frac{\Gamma(1-\frac{p}{\alpha})\Gamma(1+\frac{p}{\alpha})}{\Gamma(1-p)\Gamma(1+p)} \frac{\cos^2\frac{p\theta}{\alpha}}{\cos^2\frac{p\pi}{2}}. \tag{2.50}$$

Using (2.49), the above reduces to

$$A_pA_{-p} = \frac{\sin^2(p\pi)\Gamma(p)\Gamma(-p)}{\sin^2(\frac{p\pi}{\alpha})\Gamma(\frac{p}{\alpha})\Gamma(-\frac{p}{\alpha})}\frac{\cos^2\frac{p\theta}{\alpha}}{\cos^2\frac{p\pi}{2}}. \tag{2.51}$$

The function $\Gamma(\cdot)$ has the property,

$$\Gamma(p+1) = p\Gamma(p) \tag{2.52}$$

thus, using equations (2.49) and (2.52), the following is obtained

$$\Gamma(p)\Gamma(-p) = -\frac{\pi}{p\sin(p\pi)} \tag{2.53}$$

and

$$\Gamma(\frac{p}{\alpha})\Gamma(-\frac{p}{\alpha}) = -\frac{\alpha\pi}{p\sin(p\pi)}. \tag{2.54}$$

Taking (2.53) and (2.54) into equation (2.51) results in

$$\frac{A_pA_{-p}}{\tan\frac{p\pi}{2}} = \frac{2\cos^2\frac{p\pi}{\alpha}}{\alpha\sin\frac{p\pi}{2}}. \tag{2.55}$$

In a similar fashion, the product S_pS_{-p} can be shown to be equal to

$$S_pS_{-p}\tan\frac{p\pi}{2} = \frac{2\sin^2\frac{p\pi}{\alpha}}{\alpha\sin\frac{p\pi}{2}}. \tag{2.56}$$

Equations (2.55) and (2.56) combined lead to the following equality.

$$\text{sinc}\left(\frac{p\pi}{\alpha}\right) = \left[q\left(\frac{A_pA_{-p}}{\tan q} + S_pS_{-p}\tan q\right)\right]^{-1} \tag{2.57}$$

where $q = \frac{p\pi}{2}$.

Using the properties of Γ functions, and the first two propositions, other closed-form expressions for α, β, and γ can be derived assuming in all cases that $\delta = 0$. These FLOM estimation relations are summarized as follows.

Sinc Estimation for α: Estimate α as the solution to

$$\text{sinc}\left(\frac{p\pi}{\alpha}\right) = \left[q\left(\frac{A_pA_{-p}}{\tan q} + S_pS_{-p}\tan q\right)\right]^{-1}. \tag{2.58}$$

It is suggested in [121] that given a lower bound α_{LB} on α, a sensible range for p is $(0, \alpha_{LB}/2)$.

Ratio Estimate for δ: Given an estimate of α, estimate θ by solving

$$S_p/A_p = \tan\left(\frac{p\theta}{\alpha}\right)/\tan\left(\frac{p\pi}{2}\right). \tag{2.59}$$

Given this estimate of θ, obtain the following estimate of δ:

$$\delta = \frac{\tan(\theta)}{\tan\left(\frac{\alpha\pi}{2}\right)}. \tag{2.60}$$

FLOM Estimate for γ: Given an estimate of α, θ, solve

$$\gamma = |\cos\theta| \left(\frac{\Gamma(1-p)}{\Gamma\left(1-\frac{p}{\alpha}\right)} \frac{\cos(p\pi/2)}{\cos(p\theta/\alpha)} A_p\right)^{1/p}. \tag{2.61}$$

Note that the estimators above are all for zero-location cases, that is, $\beta = 0$. For the more general case where $\beta \neq 0$, the data must be transformed into a centered sequence by use of (2.42), then the FLOM estimation method should be applied on the parameters of the centered sequence, and finally the resulting δ and γ must be transformed by dividing by $(2 - 2^\alpha)/(2 + 2^\alpha)$ and $(2 + 2^\alpha)^{\frac{1}{\alpha}}$ respectively.

However, there are two possible problems with the FLOM method. First, since the value of a sinc function is in a finite range, when the value of the right size of (2.58) is out of this range, there is no solution for (2.58). Secondly, estimating α needs a proper value of p, which in turn depends on the value of α; in practice this can lead to errors in choosing p.

EXAMPLE 2.4

Consider the first-order modeling of the RTT time series in Figure 1.3 using the estimators of the α-stable parameters. The modeling results are shown in Table 2.1. Figure 2.10 shows histograms of the data and the pdfs associated with the parameters estimated.

Table 2.1 Estimated parameters of the distribution of the RTT time series measured between a host at the University of Delaware and hosts in Australia, Sydney, Japan, and the United Kingdom.

Parameter	Australia	Sydney	Japan	UK
α	1.0748	1.5026	1.0993	1.2180
δ	−0.3431	1	0.6733	1
γ	0.0010	7.6170×10^{-4}	0.0025	0.0014
β	0.2533	0.2359	0.2462	0.1091

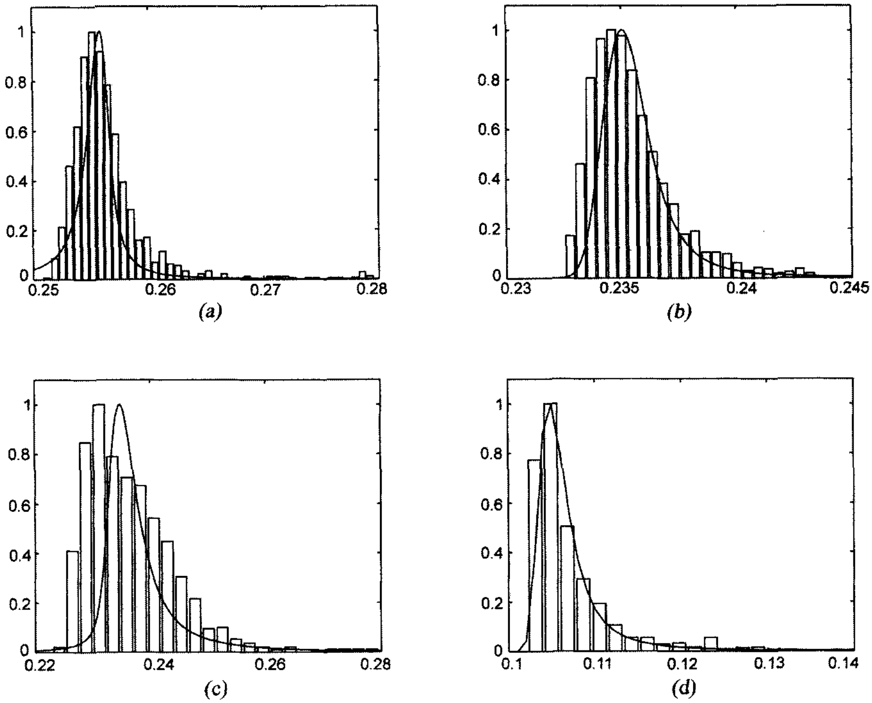


Figure 2.10 Histogram and estimated PDF of the RTT time series measured between a host at the University of Delaware and hosts in (*a*) Australia, (*b*) Sydney, (*c*) Japan, and (*d*) the United Kingdom.

■

Problems

2.1 Let $\hat{\beta}_p$ denote the L_p estimator defined by

$$\hat{\beta}_p = \arg\min_{\beta} \sum_{i=1}^{N} |x_i - \beta|^p.$$

(a) Show that when $0 < p \leq 1$, the estimator is selection-type (i.e., $\hat{\beta}_p$ is always equal to one of the input samples x_i).

(b) Define $\hat{\beta}_0 = \lim_{p \leftrightarrow 0} \hat{\beta}_p$. Prove that $\hat{\beta}_0$ is selection-type, and that it is always equal to one of the most repeated values in the sample set.

2.2 The set of well-behaved samples $\{-5, 5, -3, 3, -1, 1\}$ has been contaminated with an outlier sample of value 200.

(a) Plot the value of the L_p estimator $\hat{\beta}_p$ as a function of p, for $0 \leq p \leq 3$.

(b) Assuming that the ideal location of this distribution is $\beta = 0$, interpret the qualitative robustness of the L_p estimator as a function of p.

2.3 For $X \sim \text{Cauchy}(\gamma)$, find the mean and variance of X.

2.4 Let $X_1, \ldots, X_N$ denote a set of independent and identically distributed random variables with $X_i \sim \text{Cauchy}(1)$. Show that the sample mean

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

posses the same distribution as any of the samples X_i. What does this tell about the efficiency of $\bar{X}$ in Cauchy noise? Can we say $\bar{X}$ is robust?

2.5 Show that Gaussian distributions are stable (i.e., show that $a^2 + b^2 = c^2$, so $\alpha = 2$).

2.6 Show that Cauchy distributions are stable (i.e., show that $a + b = c$, so $\alpha = 1$).

2.7 Find the asymptotic order of the tails of:

(a) A Gaussian distribution.

(b) A Laplacian distribution.

2.8 Show that the geometric power $S_0(X)$ satisfies:

(a) $S_0(X) \geq 0$.

(b) $S_0(cX) = |c| S_0(X)$.

2.9 Find the geometric power of $X \sim \text{Uniform}(-\sigma/2, \sigma/2)$.

2.10 Let W be exponentially distributed with mean 1. Show that $W = -\ln U$ with $U \sim \text{Uniform}(0, 1)$.

2.11 Show that the expressions in equations (2.42), (2.43), and (2.44) generate centered, deskewed, and symmetrized sequences with the parameters indicated.

3

Order Statistics

The subject of order statistics deals with the statistical properties and characteristics of a set of variables that have been ordered according to magnitude. Represent the elements of an observation vector $\mathbf{X} = [X(n), X(n-1), \ldots, X(n-N+1)]^T$, as $\mathbf{X} = [X_1, X_2, \ldots X_N]^T$. If the random variables $X_1, X_2, \ldots, X_N$ are arranged in ascending order of magnitude such that

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)},$$

we denote $X_{(i)}$ as the ith-*order statistic* for $i = 1, \ldots, N$. The extremes $X_{(N)}$ and $X_{(1)}$, for instance, are useful tools in the detection of outliers. Similarly, the range $X_{(N)} - X_{(1)}$ is well known to be a quick estimator of the dispersion of a sample set.

An example to illustrate the applications of order statistics can be found in the ranking of athletes in Olympic sports. In this case, a set of N judges, generally from different nationalities, judge a particular athlete with a score bounded by a minimum assigned to a poor performance, and a maximum for a perfect score. In order to compute the overall score for a given athlete, the scores of the judges are not simply averaged. Instead, the maximum and the minimum scores given by the set of judges are discarded and the remaining scores are then averaged to provide the final score. This *trimming* of the data set is consistently done because of the possible bias of judges for a particular athlete. Since this is likely to occur in an international competition, the trimmed-average has evolved into the standard method of computing Olympic scores. This simple example shows the benefit of discarding, or discriminating against, a subset of samples from a larger data set based on the information provided by the sorted data.

Sorting the elements in the observation vector $\mathbf{X}$ constitutes a nonlinear permutation of the input vector. Consequently, even if the statistical characteristics of the input vector are exactly known, the statistical description of the sorted elements is often difficult to obtain. Simple mathematical expressions are only possible for samples which are mutually independent. Note that even in this simple case where the input samples $X_1, \ldots, X_N$ are statistically independent, the order statistics are necessarily dependent because of the ordering on the set.

The study of order statistics originated as a result of mathematical curiosity. The appearance of Sarhan and Greenberg's edited volume (1962) [171], and H. A. David's treatise on the subject (1970) [58] have changed this. Order statistics have since received considerable attention from numerous researchers. A classic and masterful survey is found in H. A. David (1981) [58]. Other important references include the work on extreme order statistics by Galambos (1978) [77], Harter's treatment in testing and estimation (1970) [96], Barnett and Lewis' (1984) [28] use of order statistics on data with outliers, and the introductory text of Arnold, Balakrishnan, and Nagaraja (1992) [16]. Parallel to the theoretical advances in the area, order statistics have also found important applications in diverse areas including life-testing and reliability, quality control, robustness studies, and signal processing. *The Handbook of Statistics Vol. 17*, edited by Balakrishnan and Rao (1998) [24], provides an encyclopedic survey of the field of order statistics and their applications.

3.1 DISTRIBUTIONS OF ORDER STATISTICS

When the variables are independent and identically distributed (i.i.d.), and when the parent distribution is continuous, the density of the rth order statistic is formed as follows. First, decompose the event that $x < X_{(r)} \leq x + dx$ into three exclusive parts: that $r-1$ of the samples X_i are less than or equal to x, that one is between x and $x + dx$, and that $N-r$ are greater than $x + dx$. Figure 3.1*a* depicts the configuration of such event. The probability that $N-r$ are greater than or equal to $x + dx$ is simply $[1 - F(x + dx)]^{N-r}$, the probability that one is between x and $x + dx$ is $f_x(x)\, dx$, and the probability that $r-1$ are less than or equal to x is $F(x)^{r-1}$. The probability corresponding to the event of having more than one sample in the interval $(x, x + dx]$ is on the order of $(dx)^2$ and is negligible as dx approaches zero. The objective is to enumerate all possible outcomes of the $X_i^{\prime}s$ such that the ordering partition is satisfied. Counting all possible enumerations of N samples in the three respective groups and using the fact that $F(x + dx) \to F(x)$ as $dx \to 0$, we can write

$$\begin{aligned} f_{(r)}(x)\, dx &= Pr\left[\, x < X_{(r)} \leq x + dx \,\right] \\ &= \frac{N!}{(r-1)!\,(N-r)!} F(x)^{r-1}\, [1 - F(x)]^{N-r}\, f_x(x)\, dx. \qquad (3.1) \end{aligned}$$

The density function of the rth order statistic, $f_{(r)}(x)$, follows directly from the above. The coefficient in the right side of (3.1) is the trinomial coefficient whose structure follows from the general multinomial coefficient as described next. Given a set of N objects, k_1 labels of type 1, k_2 labels of type 2, . . ., and k_m labels of type m and suppose that $k_1 + k_2 + \ldots + k_m = N$, the number of ways in which we may assign the labels to the N objects is given by the multinomial coefficient

$$\frac{N!}{k_1!\, k_2! \cdots k_m!}\,. \tag{3.2}$$

The trinomial coefficient in (3.1) is a special case of (3.2) with $k_1 = r - 1$, $k_2 = 1$, and $k_3 = N - r$.

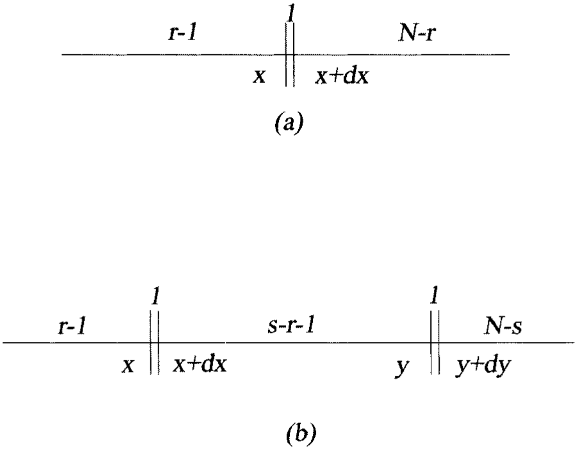


Figure 3.1 (a) The event $x < X_{(r)} \leq x + dx$ can be seen as $r - 1$ of the samples X_i are less than or equal to x, that one is between x and $x + dx$, and that $N - r$ are greater than or equal to x. (b) The event $x < X_{(r)} \leq x + dx$ and $y < X_{(s)} \leq y + dy$ can be seen as $r - 1$ of the samples X_i are less than x, that one of the samples is between x and $x + dx$, that $s - r - 1$ of the samples X_i are less than y but greater than x, that one of the samples is between y and $y + dy$, and finally that $N - s$ of the samples are greater than y.

The joint density function of the order statistics $X_{(r)}$ and $X_{(s)}$, for $1 \leq r < s \leq N$, can be found in a similar way. In this case, for $x \leq y$, the joint density is denoted as $f_{(r,s)}(x, y)$ and is obtained by decomposing the event

$$x < X_{(r)} \leq x + dx < y < X_{(s)} \leq y + dy \tag{3.3}$$

into five mutually exclusive parts: that $r - 1$ of the samples X_i are less than x, that one of the samples is between x and $x + dx$, that $s - r - 1$ of the samples X_i are less than y but greater than $x + dx$, that one of the samples is between y and $y + dy$, and finally that $N - s$ of the samples are greater than $y + dy$. The decomposition of the event in (3.3) is depicted in Figure 3.1b. The probability of occurrence for each of the five listed parts is $F(x)^{r-1}$, $f_x(x)\, dx$, $[F(y) - F(x + dx)]^{s-r-1}$, $f_x(y)\, dy$, and

$[1 - F(y+dy)]^{N-s}$. The probability corresponding to the events of having more than one sample in either of the intervals $(x, x+dx]$ and $(y, y+dy]$ is negligible as dx and dy approach zero. Using the multinomial counting principle to enumerate all possible occurrences in each part, and the fact that $F(x+dx) \sim F(x)$ and $F(y+dy) \sim F(y)$ as $dx,\ dy \to 0$ we obtain the joint density function

$$\begin{aligned} f_{(r,s)}(x,y) &= \frac{N!}{(r-1)!\,(s-r-1)!\,(N-s)!} F(x)^{r-1}\, f_x(x) \\ &\quad [F(y) - F(x)]^{s-r-1}\, f_x(y)\, [1 - F(y)]^{N-s}. \end{aligned} \tag{3.4}$$

These density functions, however, are only valid for continuous random variables, and a different approach must be taken to find the distribution of order statistics with discontinuous parent distributions. The following approach is valid for both, continuous and discontinuous distributions: let the i.i.d. variables $X_1, X_2, \ldots, X_N$ have a parent distribution function $F(x)$, the distribution function of the largest order statistic $X_{(N)}$ is

$$\begin{aligned} F_{(N)}(x) &= Pr\{X_{(N)} \leq x\} \\ &= Pr\{\text{all } X_{(i)} \leq x\} \\ &= Pr\{\text{all } X_i \leq x\} = [F(x)]^N, \end{aligned}$$

due to the independence property of the input samples. Similarly, the distribution function of the minimum sample $X_{(1)}$ is

$$\begin{aligned} F_{(1)}(x) &= Pr\{X_{(1)} \leq x\} = 1 - Pr\{X_{(1)} > x\} \\ &= 1 - Pr\{\text{all } X_i > x\} = 1 - [1 - F(x)]^N, \end{aligned}$$

since $X_{(1)}$ is less than, or equal to, all the samples in the set. The distribution function for the general case is

$$\begin{aligned} F_{(r)}(x) &= Pr\{X_{(r)} \leq x\} \\ &= Pr\{\text{at least } r \text{ of the } X_i \text{ are less than or equal to } x\} \\ &= \sum_{i=r}^{N} Pr\{\text{exactly } i \text{ of the } X_i \text{ are less than or equal to } x\} \\ &= \sum_{i=r}^{N} \binom{N}{i} [F(x)]^i [1 - F(x)]^{N-i}. \end{aligned} \tag{3.5}$$

Letting the joint distribution function of $X_{(r)}$ and $X_{(s)}$, for $1 \leq r < s \leq N$, be denoted as $F_{(r,s)}(x,y)$ then for $x < y$ we have for discrete and continuous random variables

$$\begin{aligned} F_{(r,s)}(x,y) &= Pr\{\text{at least } r \text{ of the } X_i \ \le \ x, \text{ at least } s \text{ of the } X_i \ \le \ y\ \} \\ &= \sum_{j=s}^{N}\sum_{i=r}^{j} Pr\{\text{exactly } i \text{ of } X_1, X_2 \ldots, X_n \text{ are at most } x \text{ and} \\ &\qquad \text{exactly } j \text{ of } X_1, X_2 \ldots, X_n \text{ are at most } y\} \qquad (3.6) \\ &= \sum_{j=s}^{N}\sum_{i=r}^{j} \frac{N!}{i!(j-i)!(N-j)!}[F(x)]^i[F(y)-F(x)]^{j-i}[1-F(y)]^{N-j}. \end{aligned}$$

Notice that for $x \ge y$, the ordering $X_{(r)} < x$ with $X_{(s)} \le y$, implies that $F_{(r,s)}(x,y) = F_{(s)}(y)$.

An alternate representation of the distribution function $F_{(r)}(x)$ is possible, which will prove helpful later on in the derivations of order statistics. Define the set of N samples from a uniform distribution in the closed interval $[0,1]$ as $U_1, U_2, \ldots, U_N$. The order statistics of these variates are then denoted as $U_{(1)}, U_{(2)}, \ldots, U_{(N)}$. For any distribution function $F(x)$, we define its corresponding inverse distribution function or quantile function F^{-1} as

$$F^{-1}(y) = \text{supremum } [x : F(x) \le y], \qquad (3.7)$$

for $0 < y < 1$. It is simple to show that if $X_1, \ldots, X_N$ are i.i.d. with a parent distribution $F(x)$, then the transformation $F^{-1}(U_i)$ will lead to variables with the same distribution as X_i [157]. This is written as

$$F^{-1}(U_i) \stackrel{d}{=} X_i \qquad (3.8)$$

where the symbol $\stackrel{d}{=}$ represents equality in distribution. Since cumulative distribution functions are monotonic, the smallest U_i will result in the smallest X_i, the largest U_i will result in the largest X_i, and so on. It follows that

$$F^{-1}(U_{(r)}) \stackrel{d}{=} X_{(r)}. \qquad (3.9)$$

The density function of $U_{(r)}$ follows from (3.1) as

$$f_{U_{(r)}}(u) = \frac{N!}{(r-1)!(N-r)!}u^{r-1}(1-u)^{N-r} \quad 0 < u < 1. \qquad (3.10)$$

Integrating the above we can obtain the distribution function

$$F_{U_{(r)}}(u) = \int_0^u \frac{N!}{(r-1)!(N-r)!}t^{r-1}(1-t)^{N-r}dt.$$

Using the relationship $F^{-1}(U_{(r)}) \stackrel{d}{=} X_{(r)}$, we obtain from the above the general expression

$$F_{(r)}(x) = \int_0^{F(x)} \frac{N!}{(r-1)!(N-r)!}t^{r-1}(1-t)^{N-r}dt,$$

which is an incomplete Beta function valid for any parent distribution $F(x)$ of the i.i.d. samples X_i [16].

The statistical analysis of order statistics in this section has assumed that the input samples are i.i.d. As one can expect, if the i.i.d. condition is relaxed to the case of dependent variates, the distribution function of the ordered statistics are no longer straightforward to compute. Procedures to obtain these are found in [58].

Recursive Relations for Order Statistics Distributions Distributions of order statistics can also be computed recursively, as in Boncelet (1987) [36]. No assumptions are made about the random variables. They can be discrete, continuous, mixed, i.i.d. or not.

Let $X_{(r):N}$ denote the rth order statistic out of N random variables. For first order distributions let $-\infty = t_0 < t_1 < t_2 = +\infty$ and, for second order distributions, let $-\infty = t_0 < t_1 < t_2 < t_3 = +\infty$ and let $r_1 \leq r_2$. Then, for events of order statistics:

$$\begin{aligned}
\{X_{(r):N+1} \leq t_1\} &= \{X_{(r):N} \leq t_1\}\{X_{N+1} > t_1\} && (3.11)\\
&\quad +\{X_{(r-1):N} \leq t_1\}\{X_{N+1} \leq t_1\} \\
\{X_{(r_1):N+1} \leq t_1, X_{(r_2):N+1} \leq t_2\} &= && (3.12)\\
&\quad \{X_{(r_1):N} \leq t_1, X_{(r_2):N} \leq t_2\}\{X_{N+1} > t_2\} \\
&\quad +\{X_{(r_1):N} \leq t_1, X_{(r_2-1):N} \leq t_2\}\{t_1 < X_{N+1} \leq t_2\} \\
&\quad +\{X_{(r_1-1):N} \leq t_1, X_{(r_2-1):N} \leq t_2\}\{X_{N+1} \leq t_1\}
\end{aligned}$$

In the first order case, (3.11) states that there are two ways the rth order statistic out of $N + 1$ random variables can be less or equal than t_1: one, that the $N + 1$st is larger than t_1 and the rth order statistic out of N is less or equal than t_1 and two, the $N + 1$st is less or equal than t_1 and the $r - 1$st order statistic out of N is less or equal than t_1. In the second order case, the event in question is similarly decomposed into three events.

Notice that the events on the right hand side are disjoint since the events on X_{N+1} partition the real line into nonoverlapping segments. A direct consequence of this is a recursive formula for calculating distributions for independent X_i:

$$\begin{aligned}
P(X_{(r):N+1} \leq t_1) &= P(X_{(r):N} \leq t_1)(1 - F_{N+1}(t_1)) && (3.13)\\
&\quad +P(X_{(r-1):N} \leq t_1)F_{N+1}(t_1) \\
P(X_{(r_1):N+1} \leq t_1, X_{(r_2):N+1} \leq t_2) &= \\
&\quad P(X_{(r_1):N} \leq t_1, X_{(r_2):N} \leq t_2)(1 - F_{N+1}(t_2)) && (3.14)\\
&\quad +P(X_{(r_1):N} \leq t_1, X_{(r_2-1):N} \leq t_2)(F_{N+1}(t_2) - F_{N+1}(t_1)) \\
&\quad +P(X_{(r_1-1):N} \leq t_1, X_{(r_2-1):N} \leq t_2)F_{N+1}(t_1).
\end{aligned}$$

3.2 MOMENTS OF ORDER STATISTICS

The Nth order density function provides a complete characterization of a set of N ordered samples. These distributions, however, can be difficult to obtain. Moments of order statistics, on the other hand, can be easily estimated and are often sufficient to characterize the data. The moments of order statistics are defined in the same fashion as moments of arbitrary random variables. Here we always assume that the sample size is N. The mean or expected value of the rth order statistic is denoted as $\mu_{(r)}$ and is found as

$$\begin{aligned} \mu_{(r)} &= \int_{-\infty}^{\infty} x\, f_{(r)}(x)\, dx && (3.15) \\ &= \frac{N!}{(r-1)!\,(N-r)!} \int_{-\infty}^{\infty} x\, F(x)^{r-1}[1-F(x)]^{N-r}\, f_x(x)\, dx. \end{aligned}$$

The pth raw moment of the rth-order statistic can also be defined similarly from (3.9) and (3.10) as

$$\begin{aligned} \mu_{(r)}^{(p)} &= EX_{(r)}^{p} \\ &= \frac{N!}{(r-1)!\,(N-r)!} \int_{0}^{1} \left[F^{-1}(u)\right]^{p} u^{r-1}(1-u)^{N-r}\, du, && (3.16) \end{aligned}$$

for $1 \leq r \leq N$.

Expectation of order statistic products, or *order statistic correlation*, can also be defined, for $1 \leq r \leq s \leq N$, as

$$\begin{aligned} \mu_{(r,s):N} &= E\left(X_{(r)}X_{(s)}\right) && (3.17) \\ &= [B(r, s-r, N-s+1)]^{-1} \int_{0}^{1}\int_{0}^{1} \left[F_x^{-1}(u)F^{-1}(v)u^{r-1}\right. \\ &\qquad \left.(v-u)^{s-r-1}(1-v)\right] dv\, du \end{aligned}$$

where $B(a,b,c) = \frac{(a-1)!(b-1)!(c-1)!}{(a+b+c-1)!}$. Note that (3.17) does not allude to a time shift correlation, but to the correlation of two different order-statistic variates taken from the same sample set. The statistical characteristics of the order-statistics $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ are not homogeneous, since

$$EX_{(r)} \neq EX_{(s)} \qquad (3.18)$$

for $r \neq s$, as expected since the expected value of $X_{(r)}$ should be less than the expected value of $X_{(r+1)}$. In general, the expectation of products of order statistics are not symmetric

$$E(X_{(r)}X_{(r+s)}) \neq E(X_{(r)}X_{(r-s)}). \tag{3.19}$$

This symmetry only holds in very special cases. One such case is when the parent distribution is symmetric and where $r = (N+1)/2$ such that $X_{(r)}$ is the median.

The covariance of $X_{(r)}$ and $X_{(s)}$ is written as

$$\operatorname{cov}\left[X_{(r)}X_{(s)}\right] = E\left\{\left(X_{(r)} - \mu_{(r)}\right)\left(X_{(s)} - \mu_{(s)}\right)\right\}. \tag{3.20}$$

Tukey (1958) [187], derived the nonnegative property for the covariance of order statistics: $cov[X_{(r)}X_{(s)}] \geq 0$.

3.2.1 Order Statistics From Uniform Distributions

In order to illustrate the concepts presented above, consider N samples of a standard uniform distribution with density function $f_u(u) = 1$ and distribution function $F_u(u) = u$ for $0 \leq u \leq 1$. Letting $U_{(r)}$ be the rth smallest sample, or order statistic, the density function of $U_{(r)}$ is obtained by substituting the corresponding values in (3.1) resulting in

$$f_{(r)}(u) = \frac{N!}{(r-1)!\,(N-r)!}u^{r-1}(1-u)^{N-r}, \tag{3.21}$$

also in the interval $0 \leq u \leq 1$. The distribution function follows immediately as

$$F_{(r)}(u) = \int_0^u \frac{N!}{(r-1)!\,(N-r)!}t^{r-1}(1-t)^{N-r}\,dt, \tag{3.22}$$

or alternatively using (3.5) as

$$F_{(r)}(u) = \sum_{i=r}^{N}\binom{N}{i}u^i[1-u]^{N-i}. \tag{3.23}$$

The mode of the density function can be found at $(r-1)/(N-1)$. The kth moment of $U_{(r)}$ is found from the above as

$$\begin{aligned}
\mu_{(r)}^{(k)} &= \int_0^1 u^k f_{(r)}(u)\,du \\
&= \frac{N!}{(r-1)!(N-r)!}\int_0^1 u^k u^{r-1}(1-u)^{N-r}\,du && (3.24)\\
&= B(r+k, N-r+1)/B(r, N-r+1), && (3.25)
\end{aligned}$$

where we make use of the complete beta function

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,dt \tag{3.26}$$

for $p, q > 0$. Simplifying (3.25) leads to the kth moment

$$\mu_{(r)}^{(k)} = \frac{N!\ (r+k-1)!}{(N+k)!\ (r-1)!}. \tag{3.27}$$

In particular, the first moment of the rth-order statistic can be found as

$$\mu_{(r)}^{(1)} = r/(N+1).$$

To gain an intuitive understanding of the distribution of order statistics, it is helpful to plot $f_{(r)}(u)$ in (3.21) for various values of r. For $N = 11$, Figure 3.2 depicts the density functions of the 2nd-, 3rd-, 6th- (median), 9th-, and 10th-order statistics of the samples. With the exception of the median, all other order statistics exhibit asymmetric density functions. Other characteristics of these density functions, such as their mode and shape, can be readily observed and interpreted in an intuitive fashion.

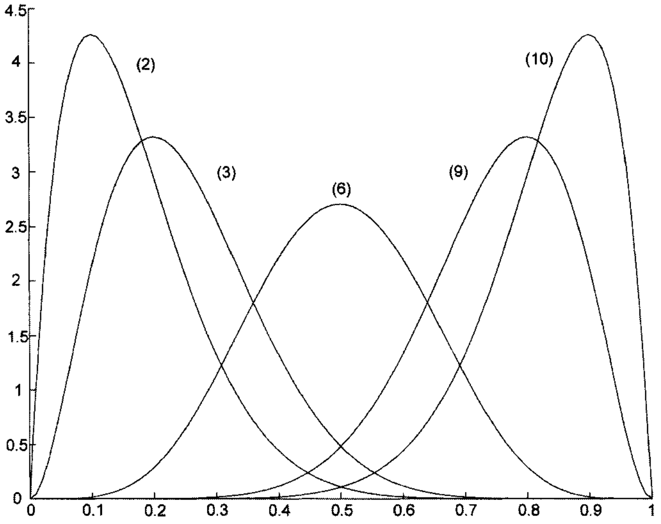


Figure 3.2 Density functions of $X_{(2)}, X_{(3)}, X_{(6)}$ (median), $X_{(9)}$, and $X_{(10)}$ for a set of eleven uniformly distributed samples.

Next consider the joint density function of $U_{(r)}$ and $U_{(s)}$ $(1 \leq r < s \leq N)$. From (3.4) we find

$$f_{(r,s)}(u,v) = \frac{N!}{(r-1)!(s-r-1)!\ (N-s)!} u^{r-1}(v-u)^{s-r-1}(1-v)^{N-s} \tag{3.28}$$

Again there are two equivalent expressions for the joint cumulative distribution function, the first is obtained integrating (3.28) and the second from Eq. (3.6)

$$\begin{aligned}
F_{(r,s)}(u,v) &= \int_0^u \int_{t_1}^v \frac{N!}{(r-1)!(s-r-1)!\,(N-s)!} \\
&\qquad \times\ t_1^{r-1}(t_2-t_1)^{s-r-1}(1-t_2)^{N-s}\,dt_2\,dt_1 \\
&= \sum_{j=s}^{N}\sum_{i=r}^{j} \frac{N!}{i!\,(j-i)!\,(N-j)!} u^i(v-u)^{j-i}\,(1-v)^{N-j}
\end{aligned}$$

for $0 \leq u < v \leq 1$. The joint density function allows the computation of the (k_r, k_s)th product moment of $(U_{(r)}, U_{(s)})$, which, after some simplifications, is found as

$$\mu_{(r,s)}^{(k_r,k_s)} = \frac{N!\ (r+k_r-1)!}{(N+k_r+k_s)!(r-1)!} \frac{(s+k_r+k_s-1)!}{(s+k_r-1)!}. \tag{3.29}$$

In particular, for $k_r = k_s = 1$, the joint moment becomes

$$\mu_{(r,s)} = \frac{r\,(s+1)}{(N+1)(N+2)}. \tag{3.30}$$

As with their marginal densities, an intuitive understanding of bivariate density functions of order statistics can be gained by plotting $f_{(r,s)}(u,v)$. Figure 3.3 depicts the bivariate density function, described in (3.28) for the 2nd- and 6th- (median) order statistics of a set of eleven uniformly distributed samples. Note how the marginal densities are satisfied as the bivariate density is integrated over each variable. Several characteristics of the bivariate density, such as the constraint that only regions where $u < v$ will have mass, can be appreciated in the plot.

3.2.2 Recurrence Relations

The computation of order-statistic moments can be difficult to obtain for observations of general random variables. In such cases, these moments must be evaluated by numerical procedures. Moments of order statistics have been given considerable importance in the statistical literature and have been numerically tabulated extensively for several distributions [58, 96]. Order-statistic moments satisfy a number of recurrence relations and identities, which can reduce the number of direct computations. Many of these relations express higher-order moments in terms of lower-order moments, thus simplifying the evaluation of higher-order moments. Since the recurrence relations between moments often involve sample sets of lower orders, it is convenient to introduce the notation $X_{(i):N}$ to represent the ith-order statistic taken from a set of N samples. Similarly, $\mu_{(i):N}$ represents the expected value of $X_{(i):N}$.

Many recursive relations for moments of order-statistics are derived from the identities

$$\sum_{i=1}^{N} X_{(i):N}^k = \sum_{i=1}^{N} X_i^k \tag{3.31}$$

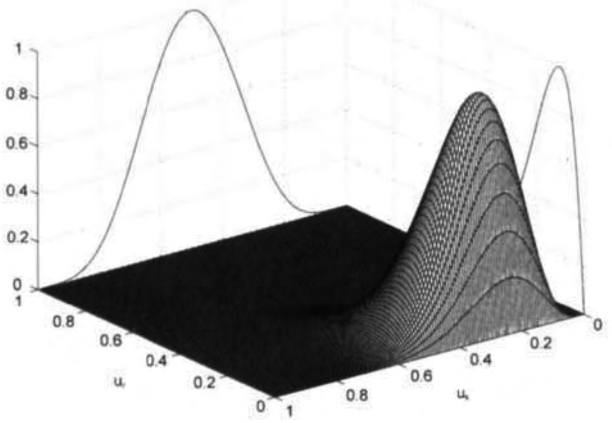


Figure 3.3 Bivariate density function of $X_{(6)}$ (median) and $X_{(2)}$ for a set of eleven uniformly distributed samples.

for $k \geq 1$, and

$$\sum_{i=1}^{N}\sum_{j=1}^{N} X_{(i):N}^{k_i} X_{(j):N}^{k_j} = \sum_{i=1}^{N}\sum_{j=1}^{N} X_i^{k_i} X_j^{k_j} \tag{3.32}$$

for $k_i,\ k_j \geq 1$, which follows from the principle that the sum of a set of samples raised to the kth power is unchanged by the order in which they are summed. Taking expectations of (3.31) leads to:

$$\sum_{i=1}^{N} \mu_{(i):N}^{(k)} = NE(X_i^k) = N\mu_{(1):1}^{(k)}$$

for $N \geq 2$ and $k \geq 1$. Similarly, from (3.32) the following is obtained:

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N} \mu_{(i,j):N}^{(k_i,k_j)} &= NE(X^{k_i+k_j}) + N(N-1)E(X^{k_i})E(X^{k_j}) \\ &= N\mu_{(1):1}^{(k_i+k_j)} + N(N-1)\mu_{(1):1}^{(k_i)}\,\mu_{(1):1}^{(k_j)}\end{aligned}$$

for $k_i, k_j \geq 1$.

These identities are simple and can be used to check the accuracy of computation of moments of order statistics. Some other useful recurrence relations are presented in the following properties.

PROPERTY 3.1 *For* $1 \leq i \leq N-1$ *and* $k \geq 1$.

$$i\mu_{(i+1):N}^{(k)} + (N-i)\mu_{(i):N}^{(k)} = N\mu_{(i):N-1}^{(k)}.$$

This property can be obtained from equation (3.16) as follows:

$$\begin{aligned}
N\mu_{(i):N-1}^{(k)} &= \frac{N!}{(i-1)!(N-i-1)!}\int_0^1 \left[F^{-1}(u)\right]^k u^{i-1}(1-u)^{N-i-1}du \\
&= \frac{N!}{(i-1)!(N-i-1)!}\int_0^1 \left[F^{-1}(u)\right]^k u^{i-1}(1-u)^{N-i-1} \\
&\qquad\qquad (u+1-u)du \\
&= \frac{N!}{(i-1)!(N-i-1)!}\left[\int_0^1 \left[F^{-1}(u)\right]^k u^{i}(1-u)^{N-i-1}\,du \right.\\
&\qquad \left. +\int_0^1 \left[F^{-1}(u)\right]^k u^{i-1}(1-u)^{N-i}du\right] \\
&= i\mu_{(i+1):N}^{(k)} + (N-i)\mu_{(i):N}^{(k)}\,.
\end{aligned}$$

Property 3.1 describes a relation known as the triangle rule [16], which allows one to compute the kth moment of a single order statistic in a sample of size N, if these moments in samples of size less than N are already available. By repeated use of the same recurrence relation, the kth moment of the remaining $N-1$ order statistics can be subsequently obtained. Hence, one could start with $\mu_{(N):N}^{(k)}$ or $\mu_{(1):N}^{(k)}$ and recursively find the moments of the smaller-or larger-order statistics.

A different recursion, published by Srikantan [180], can also be used to recursively compute single moments of order statistics by expressing the kth moment of the ith-order statistic in a sample of size N in terms of the kth moments of the largest order statistics in samples of size N and less.

PROPERTY 3.2 *For* $1 \le i \le N-1$ *and* $k \ge 1$.

$$\mu_{(i):N}^{(k)} = \sum_{j=i}^{N}(-1)^{j-i}\begin{pmatrix} N \\ j \end{pmatrix}\begin{pmatrix} j-1 \\ i-1 \end{pmatrix}\mu_{(j):j}^{(k)}\,.$$

The proof of this property is left as an exercise.

3.3 ORDER STATISTICS CONTAINING OUTLIERS

Order statistics have the characteristic that they allow us to discriminate against outlier contamination. Hence, when properly designed, statistical estimates using ordered statistics can ignore clearly inappropriate samples. In the context of robustness, it is useful to obtain the distribution functions and moments of order-statistics arising from a sample containing outliers. Here, the case where the contamination consists of a single outlier is considered. These results can be easily generalized to higher

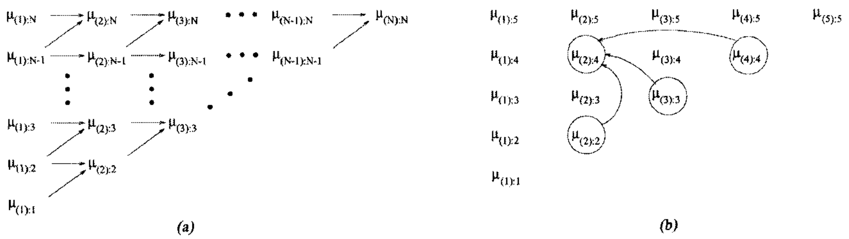


Figure 3.4 (*a*) Triangle recursion for single moments; (*b*) recurrence relation from moments of maxima of lower orders.

orders of contamination. The importance of a systematic study of order statistics from an outlier model has been demonstrated in several extensive studies [3, 59].

First, the distributions of order statistics obtained from a sample of size N when an unidentified single outlier contaminates the sample are derived. Let the N long sample set consist of $N-1$ i.i.d. variates X_i, $i = 1, \dots, N-1$, and the contaminant variable Y, which is also independent from the other samples in the sample set. Let $F(x)$ and $G(x)$ be the continuous parent distributions of X_i and Y, respectively. Furthermore, let

$$Z_{(1):N} \leq Z_{(2):N} \leq \cdots \leq Z_{(N):N} \tag{3.33}$$

be the order statistics obtained by arranging the N independent observations in increasing order of magnitude. The distribution functions of these ordered statistics are now obtained. The distribution of the maxima denoted as $H_{(N):N}(x)$ is

$$\begin{aligned} H_{(N):N}(x) &= Pr\left\{\text{all of } X_1, \dots, X_{N-1}, \text{and } Y \leq x\right\} \\ &= F(x)^{N-1}\; G(x). \end{aligned}$$

The distribution of the ith-order statistic, for $1 < i \leq N-1$, can be obtained as follows:

$$\begin{aligned} H_{(i):N}(x) &= Pr\left\{\text{ at least } i \text{ of } X_1, X_2, \dots, X_{N-1}, Y \leq x\right\} \\ &= Pr\left\{\text{exactly } i-1 \text{ of } X_1, X_2, \dots, X_{N-1} \leq x \text{ and } Y \leq x\right\} \\ &\quad + Pr\left\{\text{at least } i \text{ of } X_1, X_2, \dots, X_{N-1} \leq x\right\} \\ &= \binom{N-1}{i-1} (F(x))^{i-1}(1-F(x))^{N-i}G(x) + F_{(i):N-1}(x) \end{aligned}$$

where $F_{(i):N-1}(x)$ is the distribution of the ith-order statistic in a sample of size $N-1$ drawn from a parent distribution $F(x)$. The density function of $Z_{(i):N}$ can be obtained by differentiating the above or by direct derivation, which is left as an exercise:

$$\begin{aligned} h_{(i):N}(x) &= \frac{(N-1)!}{(i-2)!(N-i)!}(F(x))^{i-2}(1-F(x))^{N-i}G(x)f(x) \\ &+ \frac{(N-1)!}{(i-1)!(N-i)!}(F(x))^{i-1}(1-F(x))^{N-i}g(x) \\ &+ \frac{(N-1)!}{(i-1)!(N-i-1)!}(F(x))^{i-1}(1-F(x))^{N-i-1}(1-G(x))f(x) \end{aligned}$$

where the first term drops out if $i = 1$, and the last term if $N = i$.

The effect of contamination on order statistics is illustrated in Figure 3.5 depicting the densities of $Z_{(2)}$, $Z_{(6)}$ (median), and $Z_{(10)}$ for a sample set of size 11, zero-mean, double-exponential random variables. The dotted curves are the densities where no contamination exists. In the contaminated case, one of the random variables is modified such that its mean is shifted to 20. The effect of the contamination on the second-order statistic is negligible, the density of the median is only slightly affected as expected, but the effect on the 10th-order statistic, on the other hand, is severe.

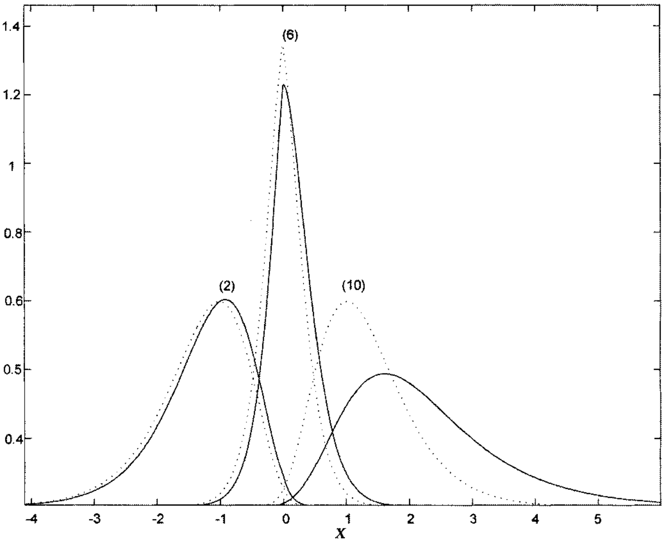


Figure 3.5 Density functions of $Z_{(2)}, Z_{(6)}$ (median), and $Z_{(10)}$ with (solid) and without contamination (dotted).

3.4 JOINT STATISTICS OF ORDERED AND NONORDERED SAMPLES

The discussion of order statistics would not be complete if the statistical relationships between the order statistics and the nonordered samples are not described. To begin, it is useful to describe the statistics of ranks. Sorting the elements $X_1, \ldots, X_N$ defines a set of N keys r_i, for $i = 1, \ldots, N$, where the rank key r_i identifies the location of X_i among the sorted set of samples $X_{(1)}, \ldots, X_{(N)}$. If the input elements to the sorter are i.i.d., each sample X_i is equally likely to be ranked first, second, or any arbitrary rank. Hence

$$P_r\left\{r_i = r\right\} = \begin{cases} \frac{1}{N} & \text{for } r = 1,\ 2, \ldots,\ N \\ 0 & \text{else.} \end{cases} \tag{3.34}$$

The expected value of each rank key is then $E\left\{r_i\right\} = (N+1)/2$. Similarly, the bivariate distribution of the two keys r_i and r_j, is given by

$$P_r\left\{r_i = r,\ r_j = s\right\} = \begin{cases} \frac{1}{N(N-1)} & \text{for } r \neq s = 1,\ 2, \ldots,\ N \\ 0 & \text{else.} \end{cases}$$

The joint distribution function of the rth order statistic $X_{(r)}$ and the ith input sample is derived next. Again, let the sample set $X_1,\ X_2, \ldots,\ X_N$ be i.i.d. with a parent distribution $F(x)$. Since the observation samples are i.i.d., the joint distribution for $X_{(r)}$ and X_i is valid for any arbitrary value of i. The joint distribution function of X_i and $X_{(r)}$ is found for $X_i \leq X_{(r)}$ as

$$\begin{aligned} F_{X_i X_{(r)}}(x,z) &= F(x) Pr\left\{X_{(r)} \leq z | X_i \leq x\right\} \\ &= F(x) Pr\left\{\text{at least } r \text{ of the } X_i's \leq z | X_i \leq x\right\}. \end{aligned}$$

Since $x \leq z$, then given that $X_i < x$ we have that the second term in the right side of the above equation is simply the probability of at least $r-1$ of the remaining $N-1$ samples $X_i < z$; thus,

$$F_{X_i X_{(r)}}(x,z) = F(x) \sum_{i=r-1}^{N-1} \binom{N-1}{i} F^i(z)(1-F(z))^{N-1-i}. \tag{3.35}$$

For the case $X_i > X_{(r)}$, the following holds:

$$\begin{aligned} F_{X_i X_{(r)}}(x,z) &= Pr\left\{X_i \leq x, X_{(r)} \leq z\right\} \\ &= [F(x) - F(z)] Pr\left\{X_{(r)} \leq z | z \leq X_i < x\right\} \\ &\quad + F(z) Pr\left\{X_{(r)} \leq z | X_i \leq z\right\}. \end{aligned}$$

These probabilities can be shown to be

$$\begin{aligned} F_{X_i X_{(r)}}(x,z) &= [F(x)-F(z)] \sum_{i=r}^{N-1} \binom{N-1}{i} F^i(z)(1-F(z))^{N-1-i} \\ &\quad + F(z) \sum_{i=r-1}^{N-1} \binom{N-1}{i} F^i(z)(1-F(z))^{N-1-i} \end{aligned}$$

for $z < x$.

The cross moments of X_i and $X_{(r)}$ can be found through the above equations, but an easier alternative method has been described in [154] as stated in the next property.

PROPERTY 3.3 *The cross moment of the rth order statistic and the nonordered sample X_i for an N i.i.d. sample set satisfies the relation*

$$E\{X_i X_{(r)}\} = \frac{1}{N} \sum_{s=1}^{N} E\{X_{(r)} X_{(s)}\}. \qquad (3.36)$$

This property follows from the relation

$$\sum_{s=1}^{N} X_{(s)} = \sum_{s=1}^{N} X_s. \qquad (3.37)$$

Substituting the above into the right hand side of (3.36) leads to

$$\begin{aligned} \sum_{s=1}^{N} E\left\{X_{(r)} X_{(s)}\right\} &= E\left\{X_{(r)} \sum_{s=1}^{N} X_{(s)}\right\} \\ &= E\left\{X_{(r)} \sum_{s=1}^{N} X_s\right\} \\ &= \sum_{s=1}^{N} E\{X_{(r)} X_s\}. \end{aligned}$$

Since all the input samples are i.i.d. then the property follows directly.

Problems

3.1 Let $X_1, \ldots, X_N$, be i.i.d. variates, X_i having a geometric density function

$$f(x) = q^x p \text{ with } q = 1 - p,$$

for $0 < p < 1$, and for $x \geq 0$. Show that $X_{(1)}$ is distributed geometrically.

3.2 For a random sample of size N from a continuous distribution whose density function is symmetrical about $x = \mu$.

(a) Show that $f_{(r)}(x)$ and $f_{(N-r+1)}(x)$ are mirror images of each other in $x = \mu$ as mirror. That is

$$f_{(r)}(\mu + x) = f_{(N-r+1)}(\mu - x).$$

(b) Generalize (a) to joint distributions of order statistics.

3.3 Let X_1, X_2, X_3 be independent and identically distributed observations taken from the density function $f(x) = 2x$ for $0 < x < 1$, and 0 elsewhere.

(a) Show that the median of the distribution is $\frac{\sqrt{2}}{2}$.

(b) What is the probability that the smallest sample in the set exceeds the median of the distribution.

3.4 Given the N marginal density functions $f_{(i)}(x)$, $1 \leq i \leq N$, of a set of i.i.d. variables, show that the average probability density function $\bar{f}(x)$ is identical to the parent density function $f(x)$. That is show

$$\bar{f}(x) = (1/N) \sum_{i=1}^{N} f_{(i)}(x) = f(x). \tag{3.38}$$

3.5 Let $X_1, X_2, \ldots, X_N$ be N i.i.d. samples with a Bernoulli parent density function such that $P_r\{X_i = 1\} = p$ and $P_r\{X_i = 0\} = 1 - p$ with $0 < p < 1$.

(a) Find $P_r\{X_{(i)} = 1\}$ and $P_r\{X_{(i)} = 0\}$.

(b) Derive the bivariate distribution function of $X_{(i)}$ and $X_{(j)}$.

(c) Find the moments $\mu_{(i)}$ and $\mu_{(i,j)}$.

3.6 Show that in odd-sized random samples from i.i.d continuous distributions, the expected value of the sample median equals the median of the parent distribution.

3.7 Show that the distribution function of the midrange $m = \frac{1}{2}(X_{(1)} + X_{(N)})$ of N i.i.d. continuous variates is

$$F(m) = N \int_{-\infty}^{m} \left[F_X(2m - x) - F_X(x)\right]^{N-1} f_X(x) dx.$$

3.8 For the geometric distribution with

$$Pr(X_i = x) = p\, q^x \quad \text{for } x \geq 0 \tag{3.39}$$

where $q = 1 - p$, show that for $1 \leq i \leq N$

$$\mu_{(i):N} = \sum_{j=N-i+1}^{N} (-1)^{j-N+i-1} \begin{pmatrix} j-1 \\ N-i \end{pmatrix} \begin{pmatrix} N \\ j \end{pmatrix} \frac{1}{(1-q^j)} \tag{3.40}$$

and

$$\mu^{(2)}_{(i):N} = \sum_{j=N-i+1}^{N} (-1)^{j-N+i-1} \begin{pmatrix} j-1 \\ N-1 \end{pmatrix} \begin{pmatrix} N \\ j \end{pmatrix} \frac{(1+q)^{j}}{(1-q^j)} . \tag{3.41}$$

3.9 Consider a set of 3 samples $\{X_1,\ X_2,\ X_3\}$. While the sample X_3 is independent and uniformly distributed in the interval $[0,1]$, the other two samples are mutually dependent with a joint density function $f(X_1, X_2) = \frac{2}{3}\delta(X_1 - 1, X_2 - 1) + \frac{1}{3}\delta(X_1 - 1, X_2)$, where $\delta(\cdot,\cdot)$ is a 2-Dimensional Dirac delta function.

(a) Find the distribution function of $X_{(3)}$.

(b) Find the distribution function of the median.

(c) Is the distribution of $X_{(1)}$ symmetric to that of $X_{(3)}$, explain.

3.10 Prove the relation in Property 3.2.

$$\mu^{(k)}_{(i):N} = \sum_{j=i}^{N} (-1)^{j-i} \begin{pmatrix} N \\ j \end{pmatrix} \begin{pmatrix} j-1 \\ i-1 \end{pmatrix} \mu^{(k)}_{(j):j} \tag{3.42}$$

Hint: From the definition of $\mu^{(k)}_{(i):N}$ we get

$$\begin{aligned} \mu^{(k)}_{(i):N} &= \frac{N!}{(i-1)!(N-i)!} \int_0^1 \left[F^{-1}(u)\right]^k u^{i-1}(1-u)^m(1-u)^{N-i-m}du \\ &= \frac{N!}{(i-1)!(N-i)!} \sum_{r=0}^{m} (-1)^r \binom{m}{r} \int_0^1 [F^{-1}(u)]^k u^{i+r-1}(1-u)^{N-i-m}du, \end{aligned}$$

which can be simplified to

$$(N-i)\mu^{(k)}_{(i):N} = \sum_{r=0}^{m} (-i)^{(r)}(n)^{(m-r)} \begin{pmatrix} m \\ r \end{pmatrix} \mu^{(k)}_{(i+r):N-m+r} \tag{3.43}$$

where $(n)^m$ denotes the terms $n(n-1)\ldots(n-m+1)$.

3.11 Consider a sequence $X_1, X_2, \ldots$ of independent and identically distributed random variables with a continuous parent distribution $F(x)$. A sample X_k is called *outstanding* if $X_k > max(X_1, X_2, \ldots, X_{k-1})$ (by definition X_1 is outstanding). Prove that $Pr\{X_k > max(X_1, X_2, \ldots, X_{k-1})\} = \frac{1}{k}$.

4 Statistical Foundations of Filtering

Filtering and parameter estimation are intimately related due to the fact that information is carried, or can be inserted, into one or more parameters of a signal at hand. In AM and FM signals, for example, the information resides in the envelope and instantaneous frequency of the modulated signals respectively. In general, information can be carried in a number of signal parameters including but not limited to the mean, variance, phase, and of course frequency. The problem then is to determine the value of the information parameter from a set of observations in some optimal fashion. If one could directly observe the value of the parameter, there would be no difficulty. In practice, however, the observation contains noise, and in this case, a statistical procedure to estimate the value of the parameter is needed.

Consider a simple example to illustrate the formulation and concepts behind parameter estimation. Suppose that a constant signal β is transmitted through a channel that adds Gaussian noise Z_i. For the sake of accuracy, several independent observations X_i are measured, from which the value of β can be inferred. A suitable model for this problem is of the form

$$X_i = \beta + Z_i \qquad i = 1, 2, \ldots, N.$$

Thus, given the sample set $X_1, X_2, \ldots, X_N$, the goal is to derive a rule for processing the observations samples that will yield a good estimate of β. It should be emphasized that the parameter β, in this formulation, is unknown but fixed — there is no randomness associated with the parameter itself. Moreover, since the samples in this example deviate about the parameter β, the estimate seeks to determine the value of the *location* parameter. Estimates of this kind are known as *location estimates*. As it will become clear later on, the location estimation problem is *key* in the formulation of the optimal filtering problem.

Several methods of estimating β are possible for the example at hand. The sample mean $\bar{X}$, given by

$$\bar{\beta}_N = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

is a natural choice. An alternative would be the sample median $\tilde{\beta}_N = \tilde{X}$ in which we order the observation samples and then select the one in the middle. We might also use a *trimmed mean* where the largest and smallest samples are first discarded and the remaining $N - 2$ samples are averaged. All of these choices are valid estimates of location. Which of these estimators, if any, is best will depend on the criterion which is selected. In this Chapter, several types of location estimates are discussed. After a short introduction to the properties of estimators, the method of *maximum-likelihood estimation* is presented with criteria for the "goodness" of an estimate. The class of *M-estimators* is discussed next, generalizing the concepts behind maximum-likelihood estimation by introducing the concept of *robust* estimation. The application of location estimators to the smoothing of signals is introduced at the end of the Chapter.

4.1 PROPERTIES OF ESTIMATORS

For any application at hand, as in our example, there can be a number of possible estimators from which one can choose. Of course, one estimator may be adequate for some applications but not for others. Describing how good an estimator is, and under which circumstances, is important. Since estimators are in essence procedures that use observations that are random variables, then the estimators themselves are random variables. The estimates, as for any random variable, can be described by a probability density function. The probability density function of the estimate $\hat{\beta}$ is denoted as $f_{\hat{\beta}}(y|\beta)$, where y is a possible value for the estimate. Since this density function can change for different estimation rules, the densities alone provide a cumbersome description. Instead, we can recourse to the statistical properties of the estimates as a mean to quantify their characteristics. The statistical properties can, in turn, be used for purposes of comparison among various estimation alternatives.

Unbiased Estimators A typical probability density $f_{\hat{\beta}}(y|\beta)$ associated with an estimate is given in Figure 4.1, where the actual value of the parameter β is shown. It would be desirable for the estimate $\hat{\beta}$ to be relatively close to the actual value of β. It follows that a good estimator will have its density function as clustered together as possible about β. If the density is not clustered or if it is clustered about some other point, it is a less good estimator. Since the mean and variance of the density are good measures of where and how clustered the density function is, a good estimator is one for which the mean of $\hat{\beta}$ is close to β and for which the variance of $\hat{\beta}$ is small.

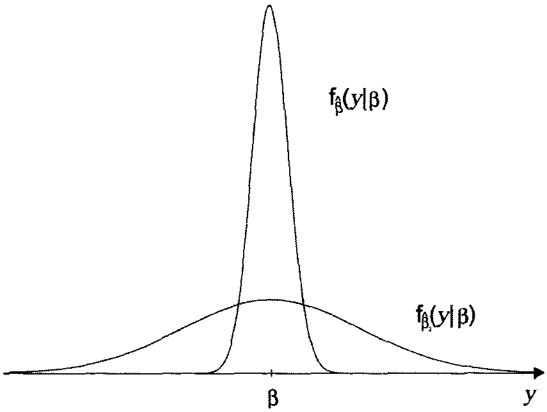


Figure 4.1 Probability density function associated with an unbiased location estimator.

In some cases, it is possible to design estimators for which the mean of $\hat{\beta}$ is always equal to the true value of β. When this desirable property is true for all values of β, the estimator is referred to as *unbiased.* Thus, the N-sample estimate of β, denoted as $\hat{\beta}_N$, is said to be *unbiased* if

$$E\{\hat{\beta}_N\} = \beta.$$

In addition, the variance of the estimate determines the precision of the estimate. If an unbiased estimate has low variance, then it will provide a more reliable estimate than other unbiased estimates with inherently larger variances. The sample mean in the previous example, is an unbiased estimate since $E\{\hat{\beta}_N\} = \beta$, with a variance that follows

$$\begin{aligned} \operatorname{var}\left\{\hat{\beta}_N\right\} &= \operatorname{var}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i\right\} \\ &= \frac{\sigma^2}{N}, \end{aligned} \tag{4.1}$$

where σ^2 is the channel noise variance. Clearly, the precision of the estimate improves as the number of observations increases.

Efficient Estimators The mean and variance of an estimate are indicators of quality. If we restrict our attention to only those estimators that are unbiased, we are in effect reducing the measure of quality to one dimension where we can define the best estimator in this class as the one that attains the minimum variance. Although at first, this may seem partially useful since we would have to search among all unbiased estimators to determine which has the lowest variance, it turns out that a lower bound on the variance of *any* unbiased estimator exists. Thus, if a given estimator is found to have a variance equal to that of the bound, the *best* estimator has been identified.

The bound is credited to Cramér and Rao [56]. Let $f(\mathbf{X};\beta)$ be the density function of the observations $\mathbf{X}$ given the value of β. For a scalar real parameter, if $\hat{\beta}$ is an unbiased estimate of β, its variance is bounded by

$$\operatorname{var}\{\hat{\beta}\} \geq \left(E\left\{\left[\frac{\partial}{\partial\beta}\ln f(\mathbf{X};\beta)\right]^2\right\}\right)^{-1} \tag{4.2}$$

provided that the partial derivative of the log likelihood function exists and is absolutely integrable. A second form of the Cramér-Rao bound can be written as

$$\operatorname{var}\{\hat{\beta}\} \geq \left(-E\left\{\frac{\partial^2}{\partial\beta^2}\ln f(\mathbf{X};\beta)\right\}\right)^{-1}, \tag{4.3}$$

being valid if the second partial derivative of the log likelihood exists and is absolutely integrable. Proofs of these bounds can be found in [32, 126]. Although there is no guarantee that an unbiased estimate exists whose variance satisfies the Cramér-Rao bound with equality, if one is found, we are certain that it is the best estimator in the sense of minimum variance and it is referred to as an *efficient estimator.*

Efficiency can also be used as a relative measure between two estimators. An estimate is said to be efficient with respect to another estimate if it has a lower variance. If this *relative efficiency* is coupled with the order of an estimate the following concept emerges: If $\hat{\beta}_N$ is unbiased and efficient with respect to $\hat{\beta}_{N-1}$ for all N, then $\hat{\beta}_N$ is said to be *consistent.*

4.2 MAXIMUM LIKELIHOOD ESTIMATION

Having a set of observation samples, a number of approaches can be taken to derive an estimate. Among these, the method of *maximum likelihood (ML)* is the most popular approach since it allows the construction of estimators even for uncommonly challenging problems. ML estimation is based on a relatively simple concept: different distributions generate different data samples and any given data sample is more likely to have come from some population than from others [99]. Conceptually, a set of observations, $X_1, X_2, \ldots, X_N$, are postulated to be values taken on by random variables assumed to follow the joint distribution function $f(X_1, X_2, \ldots, X_N; \beta)$, where β is a parameter of the distributions. The parameter β is assumed unknown but fixed, and in parameter estimation one tries to specify the best procedure to estimate the value of the parameter β from a given set of measured data.

In the method of maximum likelihood the best estimate of β is the value $\hat{\beta}_{ML}$ for which the function $f(X_1, X_2, \ldots, X_N; \beta)$ is at its maximum

$$\hat{\beta}_{ML} = \arg\max_{\beta} f(X_1, X_2, \ldots, X_N; \beta) \tag{4.4}$$

where the parameter β is variable while the observation samples $X_1, X_2, \ldots, X_N$ are fixed. The density function when viewed as a function of β, for fixed values of the observations, is known as the *likelihood function*.

The philosophy of maximum likelihood estimation is elegant and simple. Maximum likelihood estimates are also very powerful due to the notable property they enjoy that relates them to the Cramér-Rao bound. It can be shown that if an efficient estimate exists, the maximum likelihood estimate is efficient [32]. Thanks to this property, maximum likelihood estimation has evolved into one of the most popular methods of estimation.

In maximum likelihood location estimates, the parameter of interest is the *location*. Assuming independence in this model, each of the samples in the set follows some distribution

$$P(X_i \leq x) = F(x - \beta), \tag{4.5}$$

where $F(\cdot)$ corresponds to a distribution that is symmetric about 0.

Location Estimation in Gaussian Noise Assume that the observation samples $X_1, X_2, \ldots, X_N$, are i.i.d. Gaussian with a constant but unknown mean β. The maximum-likelihood estimate of location is the value $\hat{\beta}$ which maximizes the likelihood function

$$\begin{aligned} f(X_1, X_2, \ldots, X_N; \beta) &= \prod_{i=1}^{N} f(X_i - \beta) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\beta)^2/2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} e^{-\sum_{i=1}^{N}(X_i-\beta)^2/2\sigma^2}. \end{aligned} \tag{4.6}$$

The likelihood function in (4.6) can be maximized by minimizing the argument in the exponential. Thus, the maximum-likelihood estimate of location is the value $\hat{\beta}$ that minimizes the least squares sum

$$\hat{\beta}_{ML} = \arg\min_{\beta} \sum_{i=1}^{N} (X_i - \beta)^2. \tag{4.7}$$

The value that minimizes the sum, found through differentiation, results in the sample mean

$$\hat{\beta}_{ML} = \frac{1}{N} \sum_{i=1}^{N} X_i. \tag{4.8}$$

Note that the sample mean is unbiased in the assumed model since $E\{\hat{\beta}_{ML}\} = \frac{1}{N}\sum_{i=1}^{N} E\{X_i\} = \beta$. Furthermore, as a maximum-likelihood estimate, it is efficient having its variance, in (4.1), reach the Cramér-Rao bound.

Location Estimation in Generalized Gaussian Noise Now suppose that the observed data includes samples that clearly deviate from the central data cluster. The large deviations contradict a Gaussian model. The alternative is to model the deviations with a more appropriate distribution that is more flexible in capturing the characteristics of the data. One approach is to adopt the generalized Gaussian distribution. The function used to construct the maximum-likelihood estimate of location in this case is

$$\begin{aligned} f(X_1, X_2, \ldots, X_N; \beta) &= \prod_{i=1}^{N} f_k(X_i - \beta) && (4.9) \\ &= \prod_{i=1}^{N} C\, e^{-(\alpha|X_i - \beta|)^k} && (4.10) \\ &= C^N e^{-\alpha^k \sum_{i=1}^{N} |X_i - \beta|^k}, && (4.11) \end{aligned}$$

where C and α are normalizing constants and k is the fixed parameter that models the dispersion of the data. Maximizing the likelihood function is equivalent to minimizing the argument of the exponential, leading to the following estimate of location

$$\tilde{\beta}_{ML} = \arg\min_{\beta} \sum_{i=1}^{N} |X_i - \beta|^k. \qquad (4.12)$$

Some intuition can be gained by plotting the cost function in (4.12) for various values of k. Figure 4.2 depicts the different cost function characteristics obtained for $k = 2,\ 1$, and 0.5.

When the dispersion parameter is given the value 2, the model reduces to the Gaussian assumption, the cost function is quadratic, and the estimator is, as expected, equal to the sample mean. For $k < 1$, it can be shown that the cost function exhibits several local minima. Furthermore, the estimate is of *selection* type as its value will be that of one of the samples $X_1, X_2, \ldots, X_N$. These characteristics of the cost function are shown in Figure 4.2.

When the dispersion parameter is given the value 1, the model is Laplacian, the cost function is piecewise linear and continuous, and the optimal estimator minimizes the sum of absolute deviations

$$\tilde{\beta}_{ML} = \arg\min_{\beta} \sum_{i=1}^{N} |X_i - \beta|. \qquad (4.13)$$

Although not immediately seen, the solution to the above is the sample median as it is shown next.

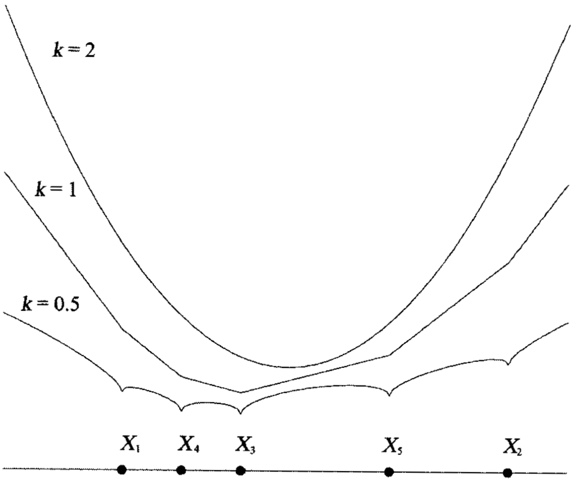


Figure 4.2 Cost functions for the observation samples $X_1 = -3, X_2 = 10, X_3 = 1, X_4 - 1, X_5 = 6$ for $k = 0.5,\ 1$, and 2.

Define the cost function being minimized in (4.13) as $L_1(\beta)$. For values of β in the interval $-\infty < \beta \leq X_{(1)}$, $L_1(\beta)$ is simplified to

$$\begin{aligned} L_1(\beta) &= \sum_{i=1}^{N} \left(X_{(i)} - \beta\right) \\ &= \sum_{i=1}^{N} X_{(i)} - N\beta. \end{aligned} \tag{4.14}$$

This, as a direct consequence that in this interval, $X_{(1)} \geq \beta$. For values of β in the range $X_{(j)} < \beta \leq X_{(j+1)}$, $L_1(\beta)$ can be written as

$$\begin{aligned} L_1(\beta) &= \sum_{i=1}^{j} \left(\beta - X_{(i)}\right) + \sum_{i=j+1}^{N} \left(X_{(i)} - \beta\right) \\ &= \left(\sum_{i=j+1}^{N} X_{(i)} - \sum_{i=1}^{j} X_{(i)} \right) - (N - 2j)\,\beta, \end{aligned} \tag{4.15}$$

for $j = 1, 2, \ldots, N-1$. Similarly, for $X_{(N)} < \beta < \infty$,

$$L_1(\beta) = -\sum_{i=1}^{N} X_{(i)} + N\beta. \tag{4.16}$$

Letting $X_{(0)} = -\infty$ and $X_{(N+1)} = \infty$, and defining $\sum_{i=m}^{n} X_{(i)} = 0$ if $m > n$, we can combine (4.14)–(4.16) into the following compactly written cost function

$$L_1(\beta) = \left(\sum_{i=j+1}^{N} X_{(i)} - \sum_{i=1}^{j} X_{(i)} \right) - (N - 2j)\,\beta, \qquad j = 0, 1, \ldots, N \tag{4.17}$$

for $\beta \in (X_{(j)}, X_{(j+1)}]$. When expressed as in (4.17), $L_1(\beta)$ is clearly piecewise linear and continuous. It starts with slope $-N$ for $-\infty < \beta \leq X_{(1)}$, and as each $X_{(j)}$ is crossed, the slope is increased by 2. At the extreme right the slope ends at N for $X_{(N)} < \beta < \infty$.

For N odd, this implies that there is an integer m, such that the slopes over the intervals $(X_{(m-1)}, X_{(m)}]$ and $(X_{(m)}, X_{(m+1)}]$, are negative and positive, respectively. From (4.17), these two conditions are satisfied if both

$$m < \frac{N}{2} \quad \text{and} \quad m > \frac{N}{2} - 1$$

hold. Both constraints are met when $m = \frac{N+1}{2}$

For N even, (4.17) implies that there is an integer m, such that the slope over the interval $(X_{(m)}, X_{(m+1)}]$ is zero. This condition is satisfied in (4.17) if

$$-(N - 2m) = 0,$$

which is possible for $m = N/2$. Thus, the maximum-likelihood estimate of location under the Laplacian model is the sample median

$$\begin{aligned} \hat{\beta}_{ML} &= \arg\min_{\beta} \sum_{i=1}^{N} |X_i - \beta| \\ &= \begin{cases} X_{\left(\frac{N+1}{2}\right)} & N \text{ odd} \\ \left(X_{\left(\frac{N}{2}\right)}, X_{\left(\frac{N}{2}+1\right)}\right] & N \text{ even} \end{cases} \\ &= \text{MEDIAN}(X_1, X_2, \ldots, X_N). \end{aligned} \tag{4.18}$$

In the case of N being even the output of the median can be any point in the interval shown above, the convention is to take the mean of the extremes $\hat{\beta}_{ML} = \frac{X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}}{2}$.

Location Estimation in Stable Noise The formulation of maximum likelihood estimation requires the knowledge of the model's closed-form density function. Among the class of symmetric stable densities, only the Gaussian ($\alpha = 2$) and Cauchy ($\alpha = 1$) distributions enjoy closed-form expressions. Thus, to formulate the non-Gaussian maximum likelihood estimation problem in a stable distribution framework, it is logical to start with the only non-Gaussian distribution for which we have a closed form expression, namely the Cauchy distribution. Although at first, this approach may seem too narrow to be effective over the broad class of stable processes, maximum-likelihood estimates under the Cauchy model can be made tunable, acquiring remarkably efficiency over the entire spectrum of stable distributions.

Given a set of i.i.d. samples $X_1, X_2, \ldots, X_N$ obeying the Cauchy distribution with scaling factor K,

$$f(x-\beta) = \frac{K}{\pi} \frac{1}{K^2 + (x-\beta)^2}, \tag{4.19}$$

the location parameter β is to be estimated from the data samples as the value $\hat{\beta}_K$, which maximizes the likelihood function

$$\hat{\beta}_K = \arg\max_{\beta} \prod_{i=1}^{N} f(X_i - \beta) = \arg\max_{\beta} \left(\frac{K}{\pi}\right)^N \prod_{i=1}^{N} \frac{1}{K^2 + (X_i - \beta)^2}. \tag{4.20}$$

This is equivalent to minimizing

$$G_K(\beta) = \prod_{i=1}^{N} [K^2 + (X_i - \beta)^2]. \tag{4.21}$$

Thus given $K > 0$, the ML location estimate is known as the sample *myriad* and is given by [82]

$$\begin{aligned} \hat{\beta}_K &= \arg\min_{\beta} \prod_{i=1}^{N} \left(K^2 + (X_i - \beta)^2\right) \qquad (4.22) \\ &= \text{MYRIAD}\{K; X_1, X_2, \ldots, X_N\}. \end{aligned}$$

Note that, unlike the sample mean or median, the definition of the sample myriad involves the free parameter K. For reasons that will become apparent shortly, we will refer to K as the *linearity parameter* of the myriad. The behavior of the myriad estimator is markedly dependent on the value of its linearity parameter K. Some intuition can be gained by plotting the cost function in (4.23) for various values of K. Figure 4.3 depicts the different cost function characteristics obtained for $K = 20, 2, 0.2$ for a sample set of size 5.

Although the definition of the sample myriad in (4.23) is straightforward, it is not intuitive at first. The following interpretations provide additional insight.

Least Logarithmic Deviation

The sample myriad minimizes $G_K(\beta)$ in (4.21), which consists of a set of products. Since the logarithm is a strictly monotonic function, the sample myriad will also minimize the expression $\log G_K(\beta)$. The sample myriad can thus be equivalently written as

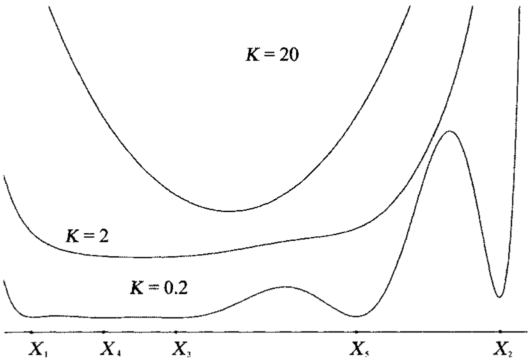


Figure 4.3 Myriad cost functions for the observation samples $X_1 = -3, X_2 = 10, X_3 = 1, X_4 - 1, X_5 = 6$ for $K = 20, 2, 0.2$.

$$\text{MYRIAD}\{K; X_1, X_2, \ldots, X_N\} = \arg\min_{\beta} \sum_{i=1}^{N} \log\left[K^2 + (X_i - \beta)^2\right]. \quad (4.23)$$

Upon observation of the above, if an observation in the set of input samples has a large magnitude such that $|X_i - \beta| \gg K$, the cost associated with this sample is approximately $\log(X_i - \beta)^2$ —the log of the square deviation. Thus, much as the sample mean and sample median respectively minimize the sum of square and absolute deviations, the sample myriad (approximately) minimizes the sum of logarithmic square deviations, referred to as the LLS criterion, in analogy to the Least Squares (LS) and Least Absolute Deviation (LAD) criteria.

Figure 4.4 illustrates the cost incurred by each sample as it deviates from the location parameter β. The cost of the sample mean (LS) is quadratic, severely penalizing large deviations. The sample median (LAD) assigns a cost that is linearly proportional to the deviation. The family of cost functions for the sample myriad assigns a penalty proportional to the logarithm of the deviation, which leads to a much milder penalization of large deviations than that imposed by the LAD and LS cost functions. The myriad cost function structure, thus, rests importance on clearly inappropriate samples.

Geometrical Interpretation

A second interpretation of the sample myriad that adds additional insight lies in its geometrical properties. First, the observations samples X_1, $X_2, \ldots,$ X_N are placed along the real line. Next, a vertical bar that runs horizontally through the real line is added as depicted in Figure 4.5. The length of the vertical bar is equal to the linearity parameter K. In this arrangement, each of the terms

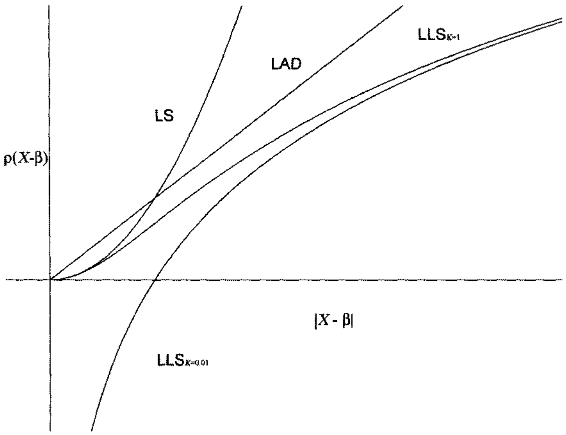


Figure 4.4 Cost functions of the mean (LS), the median (LAD), and the myriad (LLS)

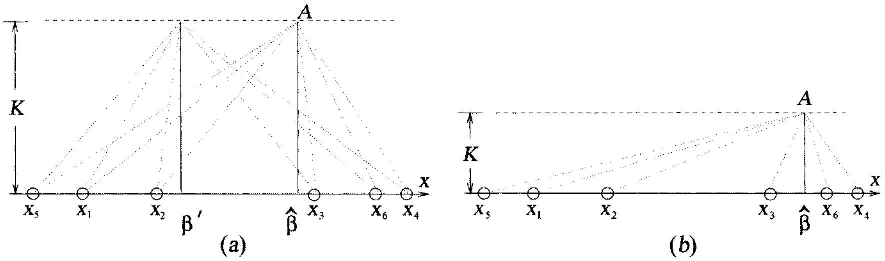


Figure 4.5 (*a*) The sample myriad, $\hat{\beta}$, minimizes the product of distances from point A to all samples. Any other value, such as $x = \beta'$, produces a higher product of distances; (*b*) the myriad as K is reduced.

$$\left(K^2 + (X_i - \beta)^2\right) \tag{4.24}$$

in (4.23), represents the distance from point A, at the end of the vertical bar, to the sample point X_i. The sample myriad, $\hat{\beta}_K$, indicates the position of the bar for which the product of distances from point A to the samples X_1, $X_2, \ldots,$ X_N is minimum. Any other value, such as $x = \beta'$, produces a higher product of distances.

If the value of K is reduced as shown in Figure 4.5*b*, the sample myriad will favor samples that are clustered together. The sample myriad has a mode-like behavior for small values of K. The term "myriad" was coined as a result of this characteristic of the estimator.

4.3 ROBUST ESTIMATION

The maximum-likelihood estimates derived so far have assumed that the form of the distribution is known. In practice, we can seldom be certain of such distributional assumptions and two types of questions arise:

(1) How sensitive are optimal estimators to the precise nature of the assumed probability model?

(2) Is it possible to construct *robust* estimators that perform well under deviations from the assumed model?

Sensitivity of Estimators To answer the first question, consider an observed data set $Z_1, Z_2, \ldots, Z_N$, and let us consider the various location estimators previously derived, namely, the mean, median, and myriad. In addition, we also consider two simple M-estimators, namely the *trimmed-mean* defined as

$$T_N(\alpha) = \frac{1}{N-2\alpha} \sum_{i=\alpha+1}^{N-\alpha} Z_{(i):N} \tag{4.25}$$

for $\alpha = 0, 1, \ldots, \lfloor N/2 \rfloor$, and the *Windsorized mean* defined as:

$$W_N(r) = \frac{1}{N} \left[\sum_{i=r+2}^{N-r-1} Z_{(i):N} + (r+1) \left[Z_{(r+1):N} + Z_{(N-r):N} \right] \right] \tag{4.26}$$

The median, is a special case of trimmed mean where $\alpha = \lfloor N/2 \rfloor$.

The effects of data contamination on these estimators is then tested. In the first set of experiments, a sample set of size 10 including one outlier is considered. The nine i.i.d. samples are distributed as $N(\mu, 1)$ and the outlier is distributed as $N(\mu + \lambda, 1)$. Table 4.1, adapted from David [58], depicts the bias of the estimation where eight different values of λ were selected.

This table clearly indicates that the mean is highly affected by the outlier. The trimming improves the robustness of the estimate. Clearly the median performs best, although it is still biased.

The expected value of the biases shown in Table 4.1 are not sufficient to compare the various estimates. The variances of the different estimators of μ are needed. These have also been tabulated in [58] and are shown on Table 4.2.

This table shows that the Windsorized mean performs better than the trimmed mean when λ is small. It also shows that, although the bias of the median is smaller, the variance is larger than the trimmed and Windsorized means. The mean is also shown to perform poorly in the MSE, except when there is no contamination.

Another useful test is to consider the contamination sample having the same mean as the other $N-1$ samples, but in this case the variance of the outlier is much larger. Hence, Table 4.3 tabulates the variance of the various estimates of μ for $N = 10$.

Table 4.1 Bias of estimators of μ for $N = 10$ when a single observation is from $N(\mu+\lambda, 1)$ and the others from $N(\mu, 1)$.

				λ				
Estimator	0.0	0.5	1.0	1.5	2.0	3.0	4.0	∞
$\bar{X}_{10}$	0.0	0.05000	0.10000	0.15000	0.20000	0.30000	0.40000	∞
$T_{10}(1)$	0.0	0.04912	0.09325	0.12870	0.15400	0.17871	0.18470	0.18563
$T_{10}(2)$	0.0	0.04869	0.09023	0.12041	0.13904	0.15311	0.15521	0.15538
Med_{10}	0.0	0.04932	0.08768	0.11381	0.12795	0.13642	0.13723	0.13726
$W_{10}(1)$	0.0	0.04938	0.09506	0.13368	0.16298	0.19407	0.20239	0.20377
$W_{10}(2)$	0.0	0.04889	0.09156	0.12389	0.14497	0.16217	0.16504	0.16530

Table 4.2 Mean squared error of various estimators of μ for $N = 10$, when a single observation is from $N(\mu + \lambda, 1)$ and the others from $N(\mu, 1)$.

				λ				
Estimator	0.0	0.5	1.0	1.5	2.0	3.0	4.0	∞
$\bar{X}_{10}$	0.10000	0.10250	0.11000	0.12250	0.14000	0.19000	0.26000	∞
$T_{10}(1)$	0.10534	0.10791	0.11471	0.12387	0.13285	0.14475	0.14865	0.14942
$T_{10}(2)$	0.11331	0.11603	0.12297	0.13132	0.13848	0.14580	0.14730	0.14745
Med_{10}	0.13833	0.14161	0.14964	0.15852	0.16524	0.17072	0.17146	0.17150
$W_{10}(1)$	0.10437	0.10693	0.11403	0.12405	0.13469	0.15039	0.15627	0.15755
$W_{10}(2)$	0.11133	0.11402	0.12106	0.12995	0.13805	0.14713	0.14926	0.14950

Table 4.3 shows that the mean is a better estimator than the median as long as the variance of the outlier is not large. The trimmed mean, however, outperforms the median regardless of the variance of the outlier. The Windsorized mean performs comparably to the trimmed mean.

These tables illustrate that by trimming the observation sample set, we can effectively increase the robustness of estimation.

M-Estimation *M-estimation* aims at answering the second question raised at the beginning of this section: Is it possible to construct estimates of location which perform adequately under deviations from distributional assumptions? According to the theory of M-estimation this is not only possible, but a well defined set of design guidelines can be followed. A brief summary of M-estimation is provided below. The interested reader can further explore the theory and applications of M-estimation in [91, 105].

Table 4.3 Variance of various estimators of μ for $N = 10$, where a single observation is from $N(\mu, \sigma^2)$ and the others from $N(\mu, 1)$.

	σ					
Estimator	0.5	1.0	2.0	3.0	4.0	∞
X_{10}	0.09250	0.10000	0.13000	0.18000	0.25000	∞
$T_{10}(1)$	0.09491	0.10534	0.12133	0.12955	0.13417	0.14942
$T_{10}(2)$	0.09953	0.11331	0.12773	0.13389	0.13717	0.14745
Med_{10}	0.11728	0.13833	0.15373	0.15953	0.16249	0.17150
$W_{10}(1)$	0.09571	0.10437	0.12215	0.13221	0.13801	0.15754
$W_{10}(2)$	0.09972	0.11133	0.12664	0.13365	0.13745	0.14950

Given a set of samples $X_1, X_2, \ldots, X_N$, an M-estimator of location is defined as the parameter $\hat{\beta}$ that minimizes a sum of the form

$$\sum_{i=1}^{N} \rho(X_i - \beta) \tag{4.27}$$

where ρ is referred to as a cost function. The behavior of the M-estimate is determined by the shape of ρ. When $\rho(x) = x^2$, for example, the associated M-estimator minimizes the sum of square deviations, which corresponds to the sample mean. For $\rho(x) = |x|$, on the other hand, the M-estimator is equivalent to the sample median. In general, if $\rho(x) = -\log f(x)$, where f is a density function, the M-estimate $\hat{\beta}$ corresponds to the maximum likelihood estimator associated with f. Accordingly, the cost function associated with the sample myriad is proportional to

$$\rho(X) = \log[k^2 + X^2]. \tag{4.28}$$

The flexibility associated with shaping $\rho(x)$ has been the key for the success of M-estimates.

Some insight into the operation of M-estimates is gained through the definition of the *influence function*. The influence function roughly measures the effect of contaminated samples on the estimates and is defined as

$$\psi(X_i - \beta) = \frac{\partial}{\partial \beta} \rho(X_i - \beta), \tag{4.29}$$

provided the derivative exists. Denoting the sample deviation $X_i - \beta$ as U_i, the influence functions for the sample mean and median are proportional to $\psi_{MEAN}(U_i) = (U_i)$ and $\psi_{MEDIAN}(U_i) = \text{sign}(U_i)$, respectively. Since the influence function of the mean is unbounded, a gross error in the observations can lead to severe distortion in the estimate. On the other hand, a similar gross error has a limited effect on the median estimate. The influence function of the sample myriad is

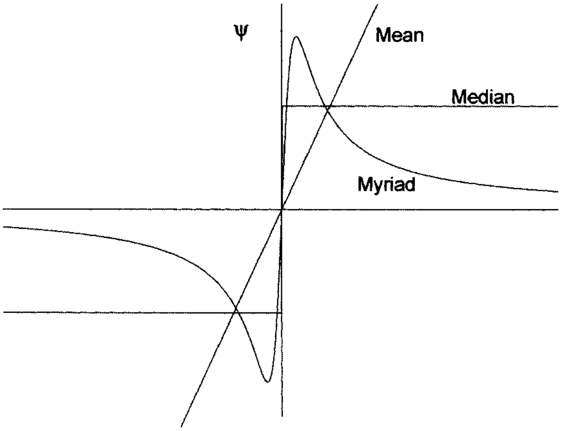


Figure 4.6 Influence functions of the mean, median and myriad

$$\psi_{MYRIAD}(U_i) = \frac{U_i}{K^2 + U_i^2}. \tag{4.30}$$

As shown in Figure 4.6, the myriad's influence function is *re-descending* reaching its maxima (minima) at $|U_i| = K$. Thus, the further away an observation sample is from the value K, the less it is considered in the estimate. Intuitively, the myriad must be more resistant to outliers than the median, and the mean is linearly sensitive to these.

Problems

4.1 Given N independent and identically distributed samples obeying the Poisson distribution:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{4.31}$$

where x can take on positive integer values, and where λ is a positive parameter to be estimated:

(a) Find the mean and variance of the random variables X_i.

(b) Derive the maximum-likelihood estimate (MLE) of λ based on a set of N observations.

(c) Is the ML estimate unbiased?

(d) Find the Cramér-Rao bound for the variance of an unbiased estimate.

(e) Find the variance of the ML estimate. Is the ML estimate efficient?

4.2 Consider N independent and identically distributed samples from a Gaussian distribution with zero mean and variance σ^2. Find the maximum likelihood estimate of σ^2 (unknown deterministic parameter). Is the estimate unbiased? Is the estimate consistent? What can you say about the ML estimate in relation to the Cramer-Rao bound.

4.3 Let X be a uniform random variable on $[\theta, \theta + 1]$, where the real-valued parameter θ is constant but unknown, and let $T(X) = \lfloor X \rfloor$= greatest integer less than or equal to X. Is $T(X)$ an unbiased estimate of θ. Hint: consider two cases: θ is an integer and θ is not an integer.

4.4 A random variable X has the uniform density

$$f(x) = 1/a \quad \text{for } 0 \leq x \leq a \tag{4.32}$$

and zero elsewhere.

(a) For independent samples of the above random variable, determine the likelihood function $f(X_1, X_2, \ldots, X_N : a)$ for $N = 1$ and $N = 2$ and sketch it. Find the maximum-likelihood estimate of the parameter a for these two cases. Find the ML estimate of the parameter a for an arbitrary number of observations N.

(b) Are the ML estimates in (a) unbiased.

(c) Is the estimate unbiased as $N \to \infty$?

4.5 Let the zero-mean random variables X and Y obey the Gaussian distribution,

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right]\right) \tag{4.33}$$

where $\rho = \frac{E[XY]}{\sigma_1\sigma_2}$ is the correlation coefficient and where $E[XY]$ is the correlation parameter. Given a set of observation pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, drawn from the joint random variables X and Y. Find the maximum likelihood estimate of the correlation parameter $E[XY]$ or of the correlation coefficient ρ.

4.6 Consider a set of N independent and identically distributed observations X_i obeying the Rayleigh density function

$$f(x) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2}. \tag{4.34}$$

(a) Find the mean and variance of the X_i variables. Note

$$\int x^m e^{ax} dx = \frac{x^m e^{ax}}{a} - \frac{m}{a}\int x^{m-1} e^{ax} dx. \tag{4.35}$$

(b) If we assume that the parameter σ^2 is unknown but constant, derive the maximum-likelihood estimate of σ^2 obtained from the N observation samples. Is the estimate unbiased?

4.7 Find the maximum-likelihood estimate of θ (unknown constant parameter) from a single observation of the variable X where

$$X = \ln \theta + N \tag{4.36}$$

where N is a noise term whose density function is unimodal with $f_N(0) > f_N(\alpha)$ for all $\alpha \neq 0$.

4.8 Consider the data set

$$X(n) = AS(n) + W(n), \qquad \text{for } n = 0, 1, \cdots, N-1, \tag{4.37}$$

where $S(n)$ is known, $W(n)$ is white Gaussian noise with known variance σ^2, and A is an unknown constant parameter.

(a) Find the maximum-likelihood estimate of A.

(b) Is the MLE unbiased?

(c) Find the variance of the MLE.

4.9 Consider N i.i.d. observations $\mathbf{X} = \{X_1, \ldots, X_N\}$ drawn from a parent distribution $F(x) = P_r(X \leq x)$. Let $\hat{F}(\mathbf{X})$ be the estimate of $F(X)$, where

$$\hat{F}(\mathbf{X}) = \frac{\text{number of } X_i's \ \leq x}{N} \tag{4.38}$$

$$\hat{F}(\mathbf{X}) = \frac{\sum_{i=1}^{N} U(x - X_i)}{N} \tag{4.39}$$

where $U(x) = 1$ if $x > 0$, and zero otherwise.

(a) Is this estimate unbiased.

(b) Prove that this estimate is the maximum-likelihood estimate. That is, let $z = \sum_{i=1}^{N} U(x - X_i)$, $\theta = F(x)$ and find $P(z/\theta)$.

This Page Intentionally Left Blank

Part II

Signal Processing with Order Statistics

This Page Intentionally Left Blank

5

Median and Weighted Median Smoothers

5.1 RUNNING MEDIAN SMOOTHERS

The running median was first suggested as a nonlinear smoother for time-series data by Tukey in 1974 [189], and it was largely popularized in signal processing by Gallagher and Wise's article in 1981 [78]. To define the running median smoother, let $\{X(\cdot)\}$ be a discrete time sequence. The running median passes a window over the sequence $\{X(\cdot)\}$ that selects, at each instant n, an odd number of consecutive samples to comprise the observation vector $\mathbf{X}(n)$. The observation window is centered at n, resulting in

$$\mathbf{X}(n) = [X(n - N_L), \dots, X(n), \dots, X(n + N_R)]^T, \tag{5.1}$$

where N_L and N_R may range in value over the nonnegative integers and $N = N_L + N_R + 1$ is the window size. In most cases, the window is symmetric about $X(n)$ and $N_L = N_R = N_1$. The median smoother operating on the input sequence $\{X(\cdot)\}$ produces the output sequence $\{Y\}$, defined at time index n as:

$$\begin{aligned} Y(n) &= \text{MEDIAN}\,[X(n - N_1), \dots, X(n), \dots, X(n + N_1)] \\ &= \text{MEDIAN}\,[X_1(n), \dots, X_N(n)] \end{aligned} \tag{5.2}$$

where $X_i(n) = X(n - N_1 - 1 + i)$ for $i = 1, 2, \dots, N$. That is, the samples in the observation window are sorted and the middle, or median, value is taken as the output. If $X_{(1)}, X_{(2)}, \dots, X_{(N)}$ are the sorted samples in the observation window, the median smoother outputs

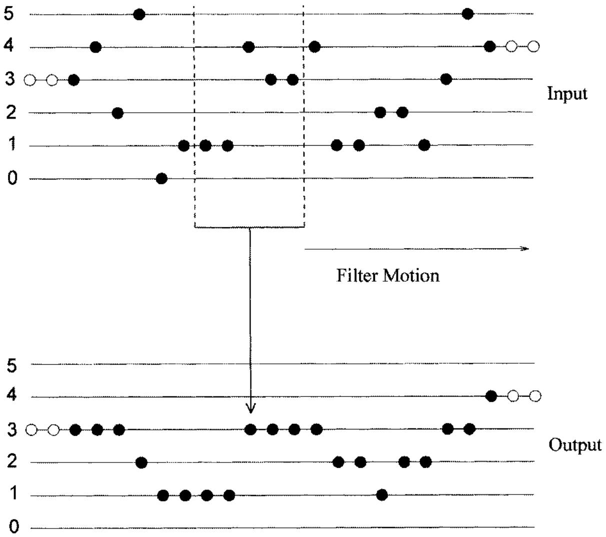


Figure 5.1 The operation of the window width 5 median smoother. o: appended points.

$$Y(n) = \begin{cases} X_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{X_{\left(\frac{N}{2}\right)}+X_{\left(\frac{N}{2}+1\right)}}{2} & \text{otherwise.} \end{cases} \tag{5.3}$$

The input sequence $\{X(\cdot)\}$ may be either finite or infinite in extent. For the finite case, the samples of $\{X(\cdot)\}$ can be indexed as $X(1),\ X(2), \ldots,\ X(L)$, where L is the length of the sequence. Because of the symmetric nature of the observation window, the window extends beyond the finite extent of the input sequence at both the beginning and end. When the window is centered at the first and last point in the signal, half of the window is empty. These end effects are generally accounted for by appending N_L samples at the beginning and N_R samples at the end of $\{X(\cdot)\}$. Although the appended samples can be arbitrarily chosen, typically these are selected so that the points appended at the beginning of the sequence have the same value as the first signal point, and the points appended at the end of the sequence all have the value of the last signal point.

To illustrate the appending of input sequences and the median smoother operation, consider the input signal $\{X(\cdot)\}$ of Figure 5.1. In this example, $\{X(\cdot)\}$ consists of 20 observations from a 6–level process, $\{X : X(n) \in \{0,\ 1, \ldots,\ 5\}, n = 1,\ 2, \ldots,\ 20\}$. The figure shows the input sequence and the resulting output sequence for a median smoother of window size 5. Note that to account for edge effects, two samples have been appended to both the beginning and end of the sequence. The median smoother output at the window location shown in the figure is

$$Y(9) \quad = \quad \text{MEDIAN}[X(7),\ X(8),\ X(9),\ X(10),\ X(11)]$$

$$= \text{MEDIAN}[\,1,\ 1,\ 4,\ 3,\ 3\,] = 3.$$

Running medians can be extended to a recursive mode by replacing the "causal" input samples in the median smoother by previously derived output samples. The output of the recursive median smoother is given by

$$\begin{aligned} Y(n) &= \text{MEDIAN}[Y(n-N_L),\ Y(n-N_L+1), \ldots, \\ &\qquad Y(n-1),\ X(n), \ldots,\ X(n+N_R)]. \end{aligned} \tag{5.4}$$

In recursive median smoothing, the center sample in the observation window is modified before the window is moved to the next position. In this manner, the output at each window location replaces the old input value at the center of the window. With the same amount of operations, recursive median smoothers have better noise attenuation capabilities than their nonrecursive counterparts [5, 8]. Alternatively, recursive median smoothers require smaller window lengths in order to attain a desired level of noise attenuation. Consequently, for the same level of noise attenuation, recursive median smoothers often yield less signal distortion.

The median operation is nonlinear. As such, the running median does not possess the superposition property and traditional impulse response analysis is not strictly applicable. The impulse response of a median smoother is, in fact, zero for all time. Consequently, alternative methods for analyzing and characterizing running medians must be employed. Broadly speaking, two types of analysis have been applied to the characterization of median smoothers: *statistical* and *deterministic*. Statistical properties examine the performance of the median smoother, through such measures as optimality and output variance, for the case of white noise time sequences. Conversely, deterministic properties examine the smoother output characteristics for specific types of commonly occurring deterministic time sequences.

5.1.1 Statistical Properties

The statistical properties of the running median can be examined through the derivation of output distributions and statistical conditions on the optimality of median estimates. This analysis generally assumes that the input to the running median is a constant signal with additive white noise. The assumption that the noise is additive and white is quite natural, and made similarly in the analysis of linear filters. The assumption that the underlying signal is a constant is certainly convenient, but more importantly, often valid. This is especially true for the types of signals median filters are most frequently applied to, such as images. Signals such as images are characterized by regions of constant value separated by sharp transitions, or edges. Thus, the statistical analysis of a constant region is valid for large portions of these commonly used signals. By calculating the output distribution of the median filter over a constant region, the noise smoothing capabilities of the median can be measured through statistics such as the filter output variance.

The calculation of statistics such as the output mean and variance from the expressions in (3.15) and (3.16) is often quite difficult. Insight into the smoothing

Table 5.1 Asymptotic output variances for the window size N mean and running median for white input samples with uniform, Gaussian, and Laplacian distributions.

Input Sample Probability Density Function	Filter Type	
	Mean	Median
Uniform $f_x(t) = \begin{cases} \frac{1}{\sqrt{12\sigma^2}} & \text{for} - \sqrt{3\sigma^2} \le t \le \sqrt{3\sigma^2} \\ 0 & \text{otherwise} \end{cases}$	$\frac{\sigma^2}{N}$	$\frac{3\sigma^2}{N+2}$
Gaussian $f_x(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$	$\frac{\sigma^2}{N}$	$\frac{\pi\sigma^2}{2N}$
Laplacian $f_x(t) = \frac{1}{\sqrt{2\sigma^2}} e^{-\frac{\sqrt{2}}{\sigma}\lvert t-\mu \rvert}$	$\frac{\sigma^2}{N}$	$\frac{\sigma^2}{2N}$

characteristics of the median filter can, however, be gained by examining the asymptotic behavior ($N \to \infty$) of these statistics, where, under some general assumptions, results can be derived. For the case of white noise input samples, the asymptotic mean, μ_{med}, and variance, σ^2_{med}, of the running median output are [126]

$$\mu_{med} = t_{0.5}, \tag{5.5}$$

and

$$\sigma^2_{med} = \frac{1}{4N(f_x(t_{0.5}))^2}, \tag{5.6}$$

where $t_{0.5}$ is the median parameter of the input samples.

Thus, the median smoother produces a consistent ($\lim_{N\to\infty} \sigma^2_{med} = 0$) and unbiased estimate of the median of the input distribution. Note that the output variance is not proportional to the input variance, but rather $1/f_x^2(t_{0.5})$. For heavy tailed noises, $1/f_x^2(t_{0.5})$ is not related to the input variance. Therefore, the variance is proportional to the impulse magnitude, not $1/f_x^2(t_{0.5})$. Thus, the output variance of the median in this case is not proportional to the input variance. This is not true for the sample mean, and further explains the more robust behavior of the median.

The variances for the sample mean and running median output are given in Table 5.1 for the uniform, Gaussian, and Laplacian input distribution cases [58]. The results hold for all N in the uniform case and are asymptotic for the Gaussian and Laplacian cases. Note that the median performs about 3 dB better than the sample mean for the Laplacian case and 2 dB worse in the Gaussian case.

Recursive median smoothers, as expected, are more efficient than their nonrecursive counterparts in attenuating noise due to the fact that half of the data points in the window of the recursive median have already been "cleaned." Consider the simplest scenario where the recursive median smoother is applied to an i.i.d. time

Table 5.2 Relative efficiency of recursive and non-recursive medians

N	$\frac{\sigma_s^2}{\sigma_r^2}$
3	1.09
5	1.39
7	1.83
9	2.40
11	3.04
13	3.73
15	4.43

series $\{X(n)\}$ described by the cumulative distribution function $F(x)$. It has been shown that the cumulative distribution function of the output of the recursive median filter $Y(n)$ with window size N, is [5, 8]

$$F_N(y) = \Pr(Y \leq y) = \frac{F(y)^{N_1} + N(1-F(y))^{N_1}F(y)^{N_1}}{F(y)^{N_1} + 2N(1-F(y))^{N_1}F(y)^{N_1} + (1-F(y))^{N_1}}, \tag{5.7}$$

where $N_1 = (N+1)/2$. The output distribution in (5.7) can be used to measure the relative efficiency between the recursive and non-recursive (standard) medians. For a window of size N and for uniformly distributed noise, the ratio σ_s^2/σ_r^2 of the nonrecursive variance estimate to the recursive variance estimate is given in Table 5.2, where the higher efficiency of the recursive median smoother is readily seen.

To further illustrate the improved noise attenuation capability of recursive medians, consider an i.i.d. input sequence, $\{X(n)\}$ consisting of a constant signal, C, embedded in additive white noise $Z(n)$. Without loss of generality, assume $C = 0$, and that the noise is symmetrically distributed. Figure 5.2*a* shows 1000 samples of the sequence $\{X(n)\}$, where the underlying distribution is double exponential (heavy tailed). Figures 5.2*b,c* show the noisy sequence after the application of a nonrecursive and a recursive median smoothers, respectively, both of window size 7. The improved noise attenuation provided by recursion is apparent in Figures 5.2*b,c*.

A phenomenon that occurs with median smoothers in impulsive noise environment is that if several impulsive noise samples are clustered together within the window, the impulses may not be removed from the signal. This phenomenon can be observed in Figures 5.2*b,c*. To quantify such events, Mallows (1980) [137] introduced the concept of *breakdown probability* as the probability of an impulse occurring at the output of the estimator, when the probability of impulses at the input is given. In essence, the breakdown probability is a measure that indicates the robustness of a particular estimator. To derive the breakdown probability of median smoothers, let us first arbitrarily select a threshold t, such that if a noise sample exceeds such level, the sample is regarded as an impulse. Let the symmetric distribution function of the noise be $F(\cdot)$, then the probability of a noise sample being an impulse (positive or negative) is $2F(-t)$. For the recursive median filter, half of the breakdown probability is given in (5.7) with $y = -t$. The breakdown probability of nonrecursive median smoothers is found through order statistics as

$$2\sum_{\ell=N_1}^{N}\binom{N}{\ell}F(-t)^{\ell}[1-F(-t)]^{N-\ell}, \tag{5.8}$$

where $N_1 = (N+1)/2$. In Figure 5.2, the threshold $|t|$ is set to 1; thus, the probability of an impulse occurring at the input is 0.24. The breakdown probability, for the non-recursive median filter, in Figure 5.2*b* is 0.011. For the recursive median filter, this probability is 0.002. Thus, on the average, for every impulse occurring at the output of the recursive median smoother in this example, there will be 5.5 impulses at the output of the nonrecursive median smoother output.

Tables 5.3 and 5.4 show the breakdown probabilities for recursive and nonrecursive median smoothers for different values of input impulse probability, $2F(-t)$, and for different window sizes. The better noise suppression characteristics of the recursive median smoothers can be seen in Figure 5.2, and in a more quantitative way in Tables 5.3 and 5.4.

Table 5.3 Breakdown probabilities for the Non-Recursive Median Smoother

Probability p	$N = 3$	$N = 5$	$N = 7$	$N = 9$	$N = 11$	$N = 13$
0.1	0.0145	0.0023	0.0003	0.00006	0.00001	0.000003
0.2	0.0560	0.0171	0.0054	0.0017	0.00059	0.0001
0.3	0.1215	0.0532	0.0242	0.0112	0.0053	0.0025
0.4	0.2080	0.1158	0.0667	0.0391	0.0233	0.0140
0.5	0.3125	0.2070	0.1411	0.0978	0.0686	0.0048
0.6	0.4320	0.3261	0.2520	0.1976	0.1564	0.1247
0.7	0.5635	0.4703	0.3997	0.3434	0.2974	0.2589

Median smoothers are primarily used to remove undesired disturbances in data, thus their statistical characterization, in terms of output distributions, would provide the required information about the median smoothers' noise attenuation power. Unfortunately, the general output distribution can seldom be put in manageable form. Unlike linear smoothers, median smoothers have well defined deterministic properties that effectively complement their set of statistical properties. In particular, root signals (also referred to invariant and fixed points) play an important role revealing the deterministic behavior of median smoothers, and in this respect the set of root signals resemble the pass band characteristics of linear frequency-selective filters.

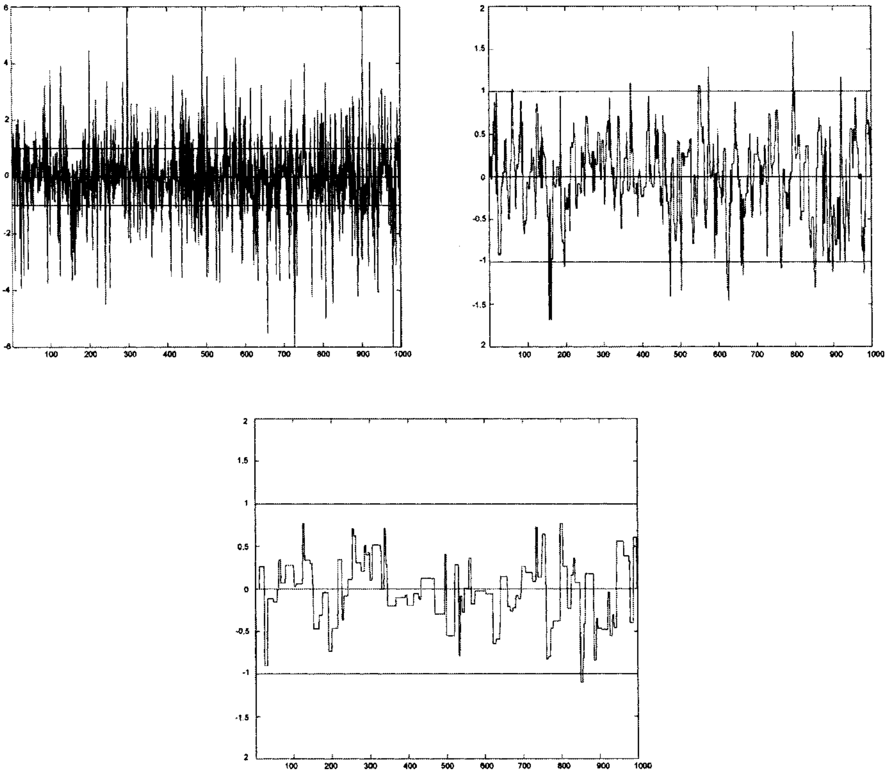


Figure 5.2 Impulse threshold $|t| = 1$: (*a*) Laplacian noisy sequence, (*b*) median smoothed sequence, and (*c*) recursive median smoothed sequence.

Table 5.4 Breakdown probabilities for the Recursive Median Smoothers

Probability p	$N = 3$	$N = 5$	$N = 7$	$N = 9$	$N = 11$	$N = 13$
0.1	0.0102	0.0007	0.00005	0.000003	0.00000002	0.00000001
0.2	0.0417	0.0066	0.0009	0.0001	0.00001	0.000001
0.3	0.0954	0.0239	0.0052	0.0010	0.0002	0.00004
0.4	0.1714	0.0604	0.0184	0.0052	0.0014	0.0003
0.5	0.2692	0.1253	0.0501	0.0187	0.0067	0.0022
0.6	0.3873	0.2285	0.1162	0.0552	0.0253	0.0113
0.7	0.5233	0.3782	0.2387	0.1417	0.0812	0.0455

5.1.2 Root Signals (Fixed Points)

Statistical properties give considerable insight into the performance of running medians. Running medians cannot, however, be sufficiently characterized through statistical properties alone. For instance, an important question not answered by the statistical properties is what type of signal, if any, is passed through a running median unaltered. Linear smoothers, when applied repeatedly to a signal, for instance, will increasingly smooth a signal. With the exception of some contrived examples, fixed points of linear smoothers are only those belonging to constant-valued sequences. On the other hand, Gallagher and Wise (1981) [78] showed that running medians have nontrivial fixed-point sequences referred to as *root signals* for reasons that will become clear shortly. The concept of root signals is important to the understanding of running medians and their effect on general signal structures. In noise smoothing, for instance, the goal is to attain maximum noise attenuation while preserving the desired signal features. An ideal situation would arise if the smoother could be tailored so that the desired signal features were invariant to the smoothing operation and only the noise would be affected. Since the median operation is nonlinear and lacks the superposition property, this idealized case is of course not possible. Nonetheless, when a signal consists of constant areas and step changes between these areas, a similar effect is achieved. Noise will be attenuated, but the signal features will remain intact. This concept is used extensively in image smoothing, where the median smoother is designed such that certain image patterns, such as lines and edges, are root signals and thus not affected by the smoothing operation [7, 147].

The definition of a root signal is quite simple: a signal is a running median root if the signal is invariant under the median smoothing operation. For simplicity assume that the window is symmetric about $X(n)$, with $N_L = N_R$ taking on the value N_1. Thus, a signal $\{X(\cdot)\}$ is a root of the window size $N = 2N_1 + 1$ median smoother if

$$X(n) = \text{MEDIAN}[X(n - N_1), \ldots, X(n), \ldots, X(n + N_1)] \qquad (5.9)$$

for all n. As an example, consider the signal shown in Figure 5.3. This signal is smoothed by three different window size running medians ($N_1 = 1, 2,$ and 3). Note that for the window size three case ($N_1 = 1$), the output is a root. That is, further smoothing of this signal with the window size three running median does not alter the signal. Notice, however, that if this same signal is smoothed with a larger window running median, the signal will be modified. Thus, the second signal (from the top) in Figure 5.3 is in the pass band, or a root, of a $N_1 = 1$ running median but outside the pass band, or not a root, of the $N_1 = 2$ and $N_1 = 3$ smoothers.

The goal of root analysis is to relate the smoothing of desired signals corrupted by noise to root and nonroot signals. If it can be shown that certain types of desired signals are in the running median root set, while noise is outside the root set, then median smoothing of a time series will preserve desired structures while altering the noise. Such a result does in fact hold and will be made clear through the following definitions and properties. First note that, as the example above illustrates, whether or not a signal is a running median root depends on the window size of the smoother in question. Clearly, all signals are roots of the window size one running median (identity). To investigate this dependence on window size, running median root signals can be characterized in terms of local signal structures, where the local signal structures are related to the window size. Such a local structure based analysis serves two purposes. First, it defines signal structures that, when properly combined, form the running median root set. Second, by relating the local structures to the window size, the effect of window size on roots is made clear. The local structure analysis of running median roots relies on the following definitions [78].

Constant Neighborhood: A region of at least $N_1 + 1$ consecutive identically valued points.

An Edge: A monotonic region between two constant neighborhoods of different value. The connecting monotonic region cannot contain any constant neighborhoods.

An Impulse: A constant neighborhood followed by at least one, but no more than N_1 points, that are then followed by another constant neighborhood having the same value as the first constant neighborhood. The two boundary points of these at most N_1 points do not have the same value as the two constant neighborhoods.

An Oscillation: A sequence of points that is not part of a constant neighborhood, an edge, or an impulse.

These definitions may now be used to develop a description of those signals that do and those that do not pass through a running median without being perturbed. In particular, Gallagher and Wise [78] developed a number of properties which characterize these signal sets for the case of finite length sequences. First, any impulse will be eliminated upon median smoothing. Secondly, a finite length signal is a running median root if it consists of constant neighborhoods and edges only. Thus,

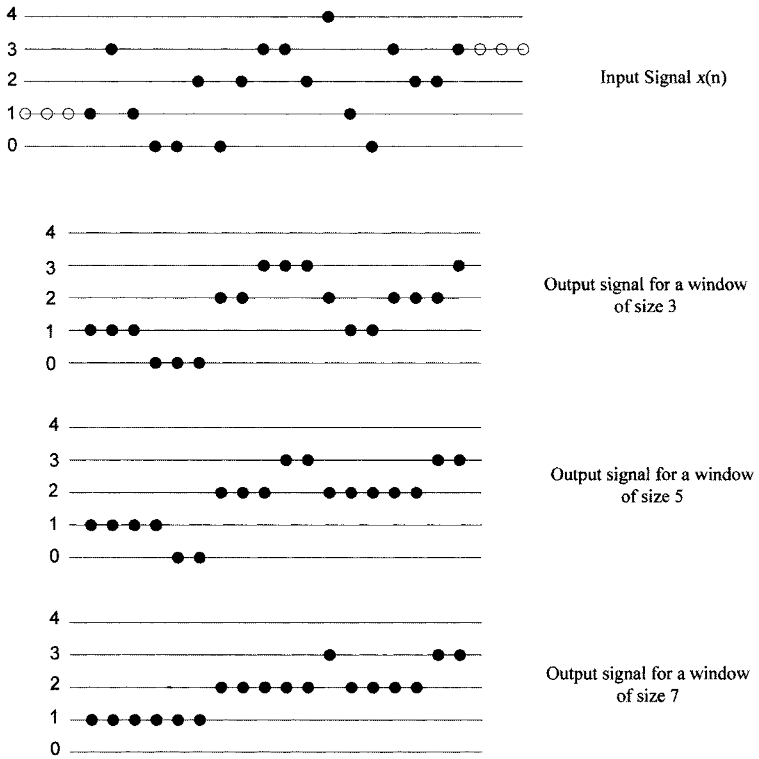


Figure 5.3 Effects of window size on a median smoothed signal. o: appended points.

if a desired signal is constructed solely of constant neighborhoods and edges, then it will not be altered by the median smoothing operation. Conversely, if observation noise consists of impulses (as defined above), it will be removed by the median smoothing operation. These running median root properties are made exact by the following.

LOMO Sequence: A sequence $\{X(\cdot)\}$ is said to be locally monotonic of length m, denoted LOMO(m), if the subsequence $X(n),\ X(n+1), \cdots,\ X(n+m-1)$ is monotonic for all $n \geq 1$.

Root Signals: Given a length L sequence to be median smoothed with a length $N = 2N_1 + 1$ window, a necessary and sufficient condition for the signal to be invariant (a root) under median smoothing is that the extended (beginning and end appended) signal be LOMO($N_1 + 2$).

Thus, the set of root signals (invariant to smoothing) of a size N running median consists solely of those signals that are formed of constant neighborhoods and edges. Note that by the definition of LOMO(m), a change of trend implies that the sequence must stay constant for at least $m - 1$ points. It follows that for a running median root signal to contain both increasing and decreasing regions, these regions must be separated by a constant neighborhood of least $N_1 + 1$ identically valued samples. It is also clear from the definition of LOMO($\cdot$) that a LOMO(m_1) sequence is also LOMO(m_2) for any two positive integers $m_1 \geq m_2$. This implies that the roots for decreasing window size running medians are nested, that is, every root of a window size M smoother is also a root of a window sized N median smoother for all $N < M$. This is formalized by:

Root Signal Set: Let S denote a set of finite length sequences and R_{N_1} be the root set of the window size $N = 2N_1+1$ running median operating on S. Then the root sets are nested such that $\ldots R_{N_1+1} \subseteq R_{N_1} \subseteq R_{N_1-1} \subseteq \ldots \subseteq R_1 \subseteq R_0 = S$.

In addition to the above description of the root signal set for running medians, it can be shown that any signal of finite length is mapped to a root signal by repeated median smoothing. This property of median filters is very significant and is called the *root convergence property*. It can be shown that the first and last points to change value on a median smoothing operation remain invariant upon additional running median passes, where repeated smoother passes consist of using the output of the prior smoothing pass for the input of an identical smoother on the current pass. This fact, in turn, indicates that any L long nonroot signal (oscillations and impulses) will become a root structure after a maximum of $(L-2)/2$ successive smoothings. This simple bound was improved in [194] where it was shown that at most

$$3\left\lceil \frac{L-2}{2(N_1+2)} \right\rceil \tag{5.10}$$

passes of the median smoother are required to reach a root. This bound is conservative in practice since in most cases root signals are obtained with much fewer smoothing passes.

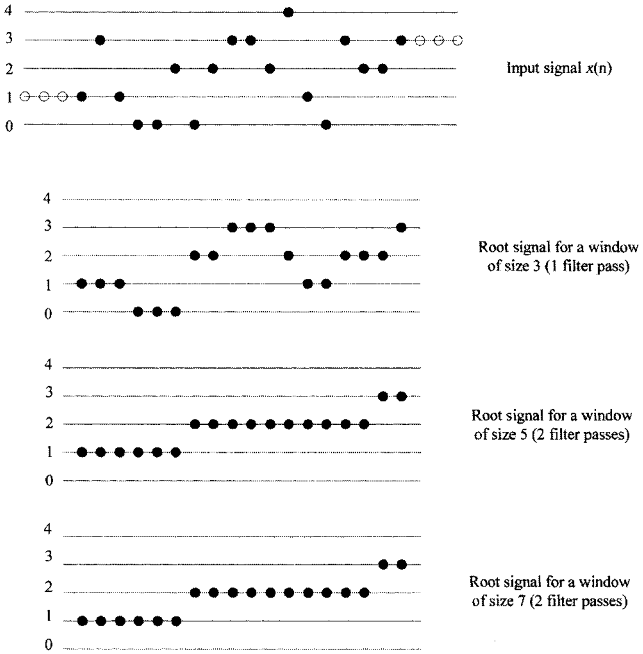


Figure 5.4 Root signals obtained by running medians of size 3, 5, and 7. o: appended points.

The running median root properties are illustrated through an example in Figure 5.4. This figure shows an original signal and the resultant root signals after multiple passes of window size 3, 5, and 7 running medians. Note that while it takes only a single pass of the window size 3 running median to obtain a root, it takes two passes for the window sizes 5 and 7 median smoothers. Clearly, the locally monotonic structure requirements of the root signals are satisfied in Figure 5.4. For the window size 3 case, the input sequence becomes LOMO(3) after a single pass of the smoother. Thus, this sequence is in the root set of the window size 3 running median, but not a root of the window size $N > 3$ running median, since it is not LOMO($N_1 + 2$) for $N_1 > 1$ ($N > 3$).

Recursive median smoothers also possess the root convergence property [5, 150]. In fact, they produce root signals after a single filter pass. For a given window size, recursive and nonrecursive median filters have the same set of root signals. A given input signal, however, may be mapped to distinct root signals by the two filters [5, 150]. Figure 5.5 illustrates this concept where a signal is mapped to different root signals by the recursive and nonrecursive median smoothers. In this case, both roots are attained in a single smoother pass.

The deterministic and statistical properties form a powerful set of tools for describing the median smoothing operation and performance. Together, they show that

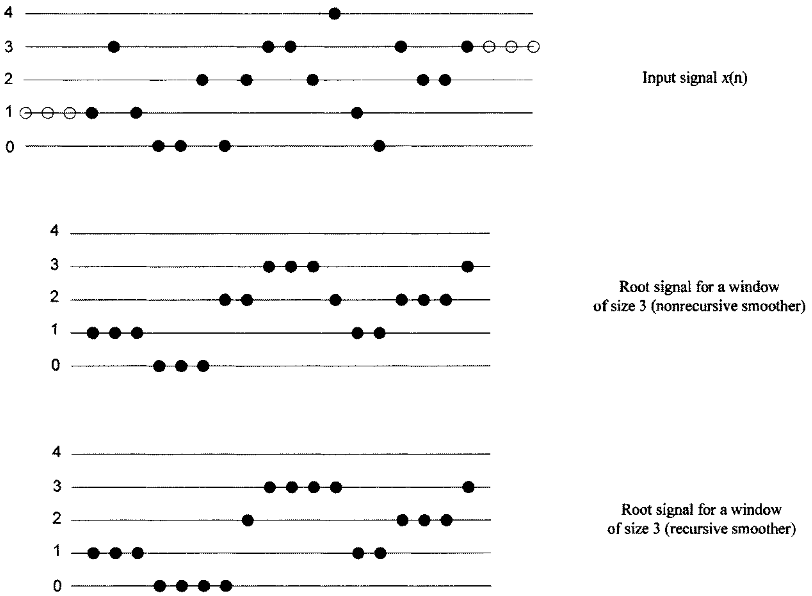


Figure 5.5 A signal and its recursive and non-recursive running median roots. o: appended points.

the median is an optimal estimator of location for Laplacian noise and that common signal structures, for example, constant neighborhoods and edges in images, are in its pass-band (root set). Moreover, impulses are removed by the smoothing operation and repeated passes of the running median always result in the signal converging to a root, where a root consists of a well defined set of structures related to the smoother's window size. Further properties of root signals can be found in Arce and Gallagher (1982) [9], Bovik (1987) [37], Wendt et al. (1986) [194], Wendt (1990) [193]. Multiscale root signal analysis was developed by Bangham (1993) [25].

MAX-MIN Representation of Medians MAX-MIN representation of medians The median has an interesting and useful representation where only minima and maxima operations are used. See Fitch (1987) [71]. This representation is useful in the software of hardware implementation of medians, but more important, it is also useful in the analysis of median operations. In addition, the max-min representation of medians provides a link between rank-order and *morphological* operators as shown in Maragos and Schafer (1987) [140]. Given the N samples X_1, $X_2, \ldots,$ X_N, and defining $m = \frac{N+1}{2}$, the median of the sample set is given by

$$X_{\left(\frac{N+1}{2}\right)} = \min\left[\max(X_1, \ldots, X_m), \ldots, \max(X_{j_1}, X_{j_2}, \ldots, X_{j_m}), \ldots, \max(X_{N-m+1}, \ldots, X_N)\right] \quad (5.11)$$

where $j_1, j_2, \ldots, j_m$ index all $C_N^m \equiv \frac{N!}{(N-m)!m!}$ combinations of N samples taken m at a time. The median of 3 samples, for instance, has the following min-max representation

$$\text{MEDIAN}(X_1,\ X_2,\ X_3) = \min\left[\max(X_1,\ X_2),\ \max(X_1,\ X_3),\ \max(X_2,\ X_3)\right]. \tag{5.12}$$

The max-min representation follows by reordering the input samples into the corresponding order-statistics $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ and indexing the resultant samples in all the possible group combinations of size m. The maximum of the first subgroup $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$ is clearly $X_{(m)}$. The maximum of the other subgroups will be greater than $X_{(m)}$ since these subgroups will include one of the elements in $X_{(m+1)}, X_{(m+2)}, \ldots, X_{(N)}$. Hence, the minimum of all these maxima will be the mth-order statistic $X_{(m)}$, that is, the median.

EXAMPLE 5.1

Consider the vector $X = [1,\ 3,\ 2,\ 5,\ 5]$, to calculate the median using the max-min representation we have:

$$\begin{aligned}\text{MEDIAN}(1,\ 3,\ 2,\ 5,\ 5) &= \min[\max(1,\ 3,\ 2),\ \max(1,\ 3,\ 5),\ \max(1,\ 3,\ 5),\\ &\qquad \max(1,\ 2,\ 5),\ \max(1,\ 2,\ 5),\ \max(1,\ 5,\ 5),\\ &\qquad \max(3,\ 2,\ 5),\ \max(3,\ 2,\ 5),\ \max(2,\ 5,\ 5)]\\ &= \min(3,\ 5,\ 5,\ 5,\ 5,\ 5,\ 5,\ 5,\ 5)\\ &= 3.\end{aligned}$$

■

5.2 WEIGHTED MEDIAN SMOOTHERS

Although the median is a robust estimator that possesses many optimality properties, the performance of running medians is limited by the fact that it is temporally blind. That is, all observation samples are treated equally regardless of their location within the observation window. This limitation is a direct result of the i.i.d. assumption made in the development of the median. A much richer class of smoothers is obtained if this assumption is relaxed to the case of independent, but not identically distributed, samples.

Statistical Foundations Although time-series samples, in general, exhibit temporal correlation, the independent but not identically distributed model can be used to synthesize the mutual correlation. This is possible by observing that the estimate

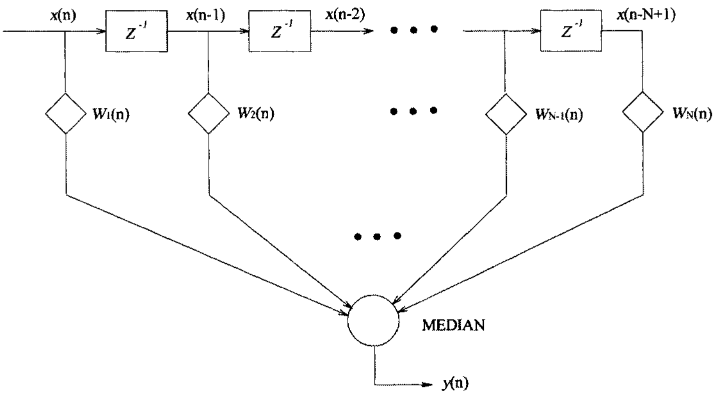


Figure 5.6 The weighted median smoothing operation.

$Y(n)$ can rely more on the sample $X(n)$ than on the other samples of the series that are further away in time. In this case, $X(n)$ is more reliable than $X(n-1)$ or $X(n+1)$, which in turn are more reliable than $X(n-2)$ or $X(n+2)$, and so on. By assigning different variances (reliabilities) to the independent but not identically distributed location estimation model, the temporal correlation used in time-series smoothing is captured. Thus, weighted median smoothers incorporate the reliability of the samples and temporal order information by weighting samples prior to rank smoothing. The WM smoothing operation can be schematically described as in Figure 5.6.

Consider again the generalized Gaussian distribution where the observation samples have a common location parameter β, but where each X_i has a (possibly) unique scale parameter σ_i. Incorporating the unique scale parameters into the ML criteria for the generalized distribution, equation (4.9), shows that, in this case, the ML estimate of location is given by the value of β minimizing

$$G_p(\beta) = \sum_{i=1}^{N} \frac{1}{\sigma_i^p} |X_i - \beta|^p. \tag{5.13}$$

In the special case of the standard Gaussian distribution ($p = 2$), the ML estimate reduces to the normalized weighted average

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(X_i - \beta)^2 = \frac{\sum_{i=1}^{N} W_i \cdot X_i}{\sum_{i=1}^{N} W_i} \tag{5.14}$$

where $W_i = 1/\sigma_i^2 > 0$. In the case of a heavier–tailed Laplacian distribution ($p = 1$), the ML estimate is realized by minimizing the sum of weighted absolute deviations

$$G_1(\beta) = \sum_{i=1}^{N} \frac{1}{\sigma_i} |X_i - \beta|. \tag{5.15}$$

where again $1/\sigma_i > 0$. Note that $G_1(\beta)$ is piecewise linear and convex for $W_i \geq 0$. The value β minimizing (5.15) is thus guaranteed to be one of the samples $X_1, X_2, \ldots, X_N$. This is the weighted median (WM), originally introduced over a hundred years ago by Edgeworth [66]. The running weighted median output is defined as

$$Y(n) = \text{MEDIAN}[W_1 \diamond X_1(n),\ W_2 \diamond X_2(n), \cdots,\ W_N \diamond X_N(n)], \qquad (5.16)$$

where $W_i > 0$ and $\diamond$ is the replication operator defined as $W_i \diamond X_i = \overbrace{X_i, \ldots, X_i}^{w_i \text{ times}}$. Weighted median smoothers were introduced in the signal processing literature by Brownigg (1984) [41] and have since received considerable attention. Note that the formulation in (5.16) requires that the weights take on nonnegative values which is consistent with the statistical interpretation of the weighted median where the weights have an inverse relationship to the variances of the respective observation samples. A simplified representation of a weighted median smoother, specified by the set of N weights, is the list of the weights separated by commas within angle brackets [202]; thus the median smoother defined in (5.16) has the representation $\langle W_1, W_2, \ldots, W_N \rangle$.

Weighted Median Computation As an example, consider the window size 5 WM smoother defined by the symmetric weight vector $\mathbf{W} = \langle 1,\ 2,\ 3,\ 2,\ 1 \rangle$. For the observation $\mathbf{X}(n) = [12,\ 6,\ 4,\ 1,\ 9]$, the weighted median smoother output is found as

$$\begin{aligned} Y(n) &= \text{MEDIAN}\left[\, 1\diamond 12,\ 2\diamond 6,\ 3\diamond 4,\ 2\diamond 1,\ 1\diamond 9 \,\right] \\ &= \text{MEDIAN}\left[\, 12,\ 6,\ 6,\ 4,\ 4,\ 4,\ 1,\ 1,\ 9 \,\right] \\ &= \text{MEDIAN}\left[\, 1,\ 1,\ 4,\ 4,\ \underline{4},\ 6,\ 6,\ 9,\ 12 \,\right] \\ &= 4 \end{aligned} \qquad (5.17)$$

where the median value is underlined in equation (5.17). The large weighting on the center input sample results in this sample being taken as the output. As a comparison, the standard median output for the given input is $Y(n) = 6$.

In general, the WM can be computed without replicating the sample data according to the corresponding weights, as this increases the computational complexity. A more efficient method to find the WM is shown next, which not only is attractive from a computational perspective but it also admits positive real–valued weights:

(1) Calculate the threshold $W_0 = \frac{1}{2}\sum_{i=1}^{N} W_i$;

(2) Sort the samples in the observation vector $\mathbf{X}(n)$;

(3) Sum the concomitant weights[1] of the sorted samples beginning with the maximum sample and continuing down in order;

(4) The output is the sample whose weight causes the sum to become $\geq W_0$.

The validity of this method can be supported as follows. By definition, the output of the WM smoother is the value of β minimizing (5.15). Suppose initially that $\beta \geq X_{(N)}$. (5.15) can be rewritten as:

$$\begin{aligned} G_1(\beta) &= \sum_{i=1}^{N} W_{[i]} \left(X_{(i)} - \beta\right) \\ &= \left(\sum_{i=1}^{N} W_{[i]}\right)\beta - \sum_{i=1}^{N} W_{[i]} X_{(i)}, \end{aligned} \tag{5.18}$$

which is the equation of a straight line with slope $m_N = \sum_{i=1}^{N} W_{[i]} \geq 0$. Now suppose that $X_{(N-1)} \leq \beta < X_{(N)}$. (5.15) is now equal to:

$$\begin{aligned} G_1(\beta) &= \sum_{i=1}^{N-1} W_{[i]} \left(X_{(i)} - \beta\right) + W_{[N]} \left(\beta - X_{(N)}\right) \\ &= \left(\sum_{i=1}^{N-1} W_{[i]} - W_{[N]}\right)\beta - \sum_{i=1}^{N-1} W_{[i]} X_{(i)} + W_{[N]} X_{(N)}. \end{aligned} \tag{5.19}$$

This time the slope of the line is $m_{N-1} = \sum_{i=1}^{N-1} W_{[i]} - W_{[N]} \leq m_N$, since all the weights are positive. If this procedure is repeated for values of β in intervals lying between the order statistics, the slope of the lines in each interval decreases and so will the value of the cost function (5.15), until the slope reaches a negative value. The value of the cost function at this point will increase. The minimum is then reached when this change of sign in the slope occurs. Suppose the minimum (i.e., the weighted median) is the Mth-order statistic. The slopes of the cost function in the intervals before and after $X_{(M)}$ are given by:

$$m_M = \sum_{i=1}^{M} W_{[i]} - \sum_{i=M+1}^{N} W_{[i]} > 0 \tag{5.20}$$

$$m_{M-1} = \sum_{i=1}^{M-1} W_{[i]} - \sum_{i=M}^{N} W_{[i]} \leq 0. \tag{5.21}$$

[1] Represent the input samples and their corresponding weights as pairs of the form (X_i, W_i). If the pairs are ordered by their X variates, then the value of W associated with $X_{(m)}$, denoted by $W_{[m]}$, is referred to as the *concomitant of the mth order statistic* [58].