OPTIMUM SIGNAL PROCESSING:
An Introduction

SOPHOCLES J. ORFANIDIS

OPTIMUM SIGNAL PROCESSING: An Introduction

SOPHOCLES J. ORFANIDIS
Rutgers University

MACMILLAN PUBLISHING COMPANY
A Division of Macmillan, Inc.
NEW YORK

Collier Macmillan Publishers
LONDON

Macmillan Publishing Company
866 Third Avenue, New York, NY 10022

Collier Macmillan Canada, Inc.

Printed in the United States of America

printing number

1 2 3 4 5 6 7 8 9 10

Library of Congress Cataloging in Publication Data

Orfanidis, Sophocles J.
Optimum signal processing.

Bibliography: p.
Includes index.
1. Signal processing. I. Title.
TK5102.5.0246 1984 621.38'043 83-24411
ISBN 0-02-949860-0

Preface

THE PURPOSE OF THIS BOOK is to provide an introduction to signal processing methods that are based on optimum Wiener filtering and least-squares estimation concepts. Such methods have a remarkably broad range of applications, ranging from the analysis and synthesis of speech, data compression, image processing and modeling, channel equalization and echo cancellation in digital data transmission, geophysical signal processing in oil exploration, linear predictive analysis of EEG signals, modern methods of high-resolution spectrum estimation, and superresolution array processing, to adaptive signal processing for sonar, radar, system identification, and adaptive control applications. The structure of the book is to present the Wiener filtering concept as the basic unifying theme that ties together the various signal processing algorithms and techniques currently used in the above applications.

The book is based on lecture notes for a second-semester graduate-level course on advanced topics in digital signal processing I have taught at Rutgers University since 1979. The book is primarily addressed to beginning graduate students in electrical engineering, but it may also be used as a reference by practicing engineers who want a concise introduction to the subject. The prerequisites for using the book are an introductory course on digital signal processing, such as on the level of Oppenheim and Schafer's book, and some familiarity with probability and random signal concepts, such as on the level of Papoulis' book.

Chapter 1 sets many of the objectives of the book and serves both as a review of probability and random signals and as an introduction to some of the basic concepts upon which the rest of the text is built. These are the concept of correlation canceling and its connection to linear mean-squared estimation, and the concept of Gram-Schmidt orthogonalization of random variables and its connection to linear prediction and signal modeling. After a brief review of some pertinent material on random signals, such as autocorrelations, power spectra, and the periodogram and its improvements, we discuss parametric signal models in which the random signal is modeled as the output of a linear system driven

by white noise and present an overview of the uses of such models in signal analysis and synthesis, spectrum estimation, signal classification, and data compression applications. A first-order autoregressive model is used to illustrate many of these ideas and to motivate some practical methods of extracting the model parameters from actual data.

Chapter 2 is also introductory, and its purpose is to present a number of straightforward applications and simulation examples that illustrate the practical usage of random signal concepts. The selected topics include simple designs for signal enhancement filters, quantization noise in digital filters, and an introduction to linear prediction based on the finite past. The last two topics are then merged into an introductory discussion of data compression by DPCM methods.

Chapter 3 introduces the concept of minimal phase signals and filters and its role in the making of parametric signal models via spectral factorization. These methods are used in Chapter 4 for the solution of the Wiener filtering problem.

The basic concept of the Wiener filter as an optimum filter for estimating one signal from another is developed in Chapter 4. The Wiener filter is also viewed as a correlation canceler and as an optimal signal separator. We consider both the stationary and nonstationary Wiener filters, as well as the more practical FIR Wiener filter. While discussing a simple first-order Wiener filter example, we take the opportunity to introduce some elementary Kalman filter concepts. We demonstrate how the steady-state Kalman filter is equivalent to the Wiener filter and how its solution may be obtained from the steady-state algebraic Riccati equation which affects the spectral factorization required in the Wiener case. We also show how the Kalman filter may be thought of as the whitening filter of the observation signal and discuss its connection to the Gram-Schmidt orthogonalization and parametric signal models of Chapter 1. This chapter is mainly theoretical in character. Practical implementations and applications of Wiener filters are discussed in Chapter 5 using block-processing methods and in Chapter 6 using real-time adaptive processing techniques.

Chapter 5 begins with a discussion of the full linear prediction problem and its connection to signal modeling and continues with the problem of linear prediction based on the finite past and its efficient solution via the Levinson recursion. We discuss the analysis and synthesis lattice filters of linear prediction, as well as the lattice realizations of more general Wiener filters that are based on the orthogonality property of the backward prediction errors. The autocorrelation, covariance, and Burg's methods of linear predictive analysis are presented, and their application to speech analysis and synthesis and to spectrum estimation is discussed. The problem of estimating the frequencies of multiple sinusoids in noise and the problem of resolving the directions of point-source emitters by spatial array processing are discussed. Four approaches to these problems are presented, namely, the classical method based on the windowed autocorrelation, the maximum entropy method based on linear prediction, Capon's maximum likelihood method, and eigenvector-based methods. We also

discuss the problem of wave propagation in layered media and its connection to linear prediction, and present the dynamic predictive deconvolution procedure for deconvolving the multiple reverberation effects of a layered structure from the knowledge of its reflection or transmission response. The chapter ends with a discussion of a least-squares reformulation of the Wiener filtering problem that can be used in the design of waveshaping and spiking filters for deconvolution applications.

Real-time adaptive implementations of Wiener filters are discussed in Chapter 6. The basic operation of an adaptive filter is explained by means of the simplest possible filter, namely, the correlation canceler loop, which forms the elementary building block of higher order adaptive filters. The Widrow-Hoff LMS adaptation algorithm and its convergence properties are discussed next. Several applications of adaptive filters are presented, such as adaptive noise canceling, adaptive channel equalization and echo cancellation, adaptive signal separation and the adaptive line enhancer, adaptive spectrum estimation based on linear prediction, and adaptive array processing. We also discuss some recent developments, such as the adaptive implementation of Pisarenko's method of harmonic retrieval, and two alternative adaptation algorithms that offer very fast speed of convergence, namely, recursive least-squares, and gradient lattice adaptive filters.

The subject of Wiener filtering and linear estimation is vast. The selection of material in this book reflects my preferences and views on what should be included in an introductory course on this subject. The emphasis thorughout the book is on the signal processing procedures that grow out of the fundamental concept of Wiener filtering. An important ingredient of the book is the inclusion of several computer experiments and assignments that demonstrate the successes and limitations of the various signal processing algorithms that are discussed. A set of FORTRAN 77 subroutines, designed to be used as a library, has been included in an appendix.

I would like to thank my colleagues Professors T. G. Marshall and P. Sannuti for their support. I am greatly indebted to Professor R. Peskin for making available his graphics system on which many of the simulation examples were run and to my graduate student Ms. E. M. Vail for her invaluable help in producing most of the computer graphs. Most of all, I would like to thank my wife Monica, without whose love and affection this book could not have been written.

SOPHOCLES J. ORFANIDIS

Contents

1

Random Signals

1.1 *Probability Density, Mean, Variance*

In this section, we present a short review of probability concepts. It is assumed that the student has had a course on the subject on the level of Papoulis' book [1].

Let x be a *random variable* having *probability density* $p(x)$. Its *mean, variance,* and *second moment* are defined as the expectation values

$$m = E[x] = \int_{-\infty}^{\infty} x\,p(x)\,dx = \text{mean}$$

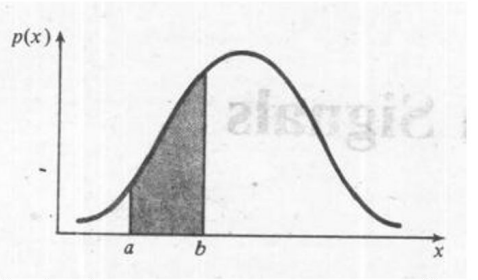
$$\sigma^2 = E[(x - m)^2] = \int_{-\infty}^{\infty} (x - m)^2\,p(x)\,dx = \text{variance}$$

$$E[x^2] = \int_{-\infty}^{\infty} x^2\,p(x)\,dx = \text{second moment}$$

These quantities are known as *second-order statistics* of the random variable x. Their importance is linked with the fact that most optimal filter design criteria require knowledge only of the second-order statistics and do not require more detailed knowledge, such as of probability densities. It is of primary importance, then, to be able to extract such quantities from the actual measured data.

The probability that the random variable x will assume a value within an interval of values $[a,b]$ is given by

$$\text{Prob}[a \leq x \leq b] = \int_a^b p(x)\,dx = \text{shaded area}$$



The probability density is always normalized to unity by

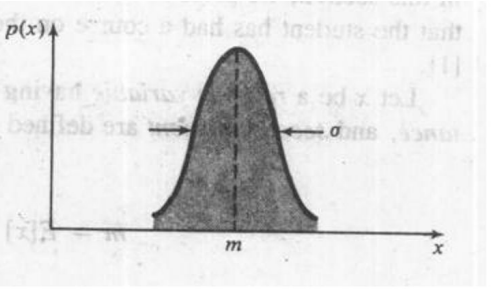
$$\int_{-\infty}^{\infty} p(x)\,dx = 1$$

which states that the probability of x taking a value somewhere within its range of variation is unity; that is, certainty. This property also implies

$$\sigma^2 = E[(x - m)^2] = E[x^2] - (E[x])^2 = E[x^2] - m^2$$

Example 1.1.1: Gaussian distribution

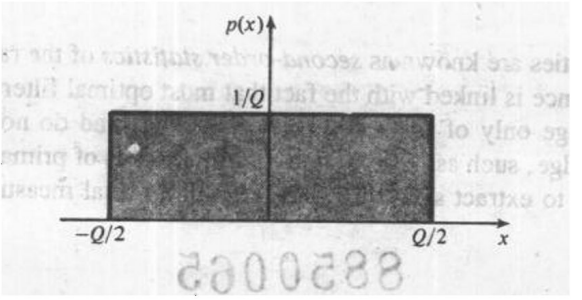
$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp[-(x - m)^2/2\sigma^2]$$



Example 1.1.2: Uniform distribution

$$p(x) = \begin{cases} 1/Q, & \text{for } -Q/2 \leq x \leq Q/2 \\ 0, & \text{otherwise} \end{cases}$$

Its variance is $\sigma^2 = Q^2/12$.



Both the gaussian and the uniform distributions will prove to be important examples. In typical signal processing problems of designing filters to remove or separate noise from signal, it is often assumed that the noise interference is gaussian. This assumption is justified on the grounds of the *central limit theorem*, provided that the noise arises from many different noise sources acting independently of each other.

The uniform distribution is also important. In digital signal processing applications, the quantization error arising from the signal quantization in the A/D converters, or the roundoff error arising from the finite accuracy of the internal arithmetic operations in digital filters, can often be assumed to be uniformly distributed.

Every computer provides system routines for the generation of random numbers. For example, the routines RANDU and GAUSS of the IBM Scientific Subroutine Package generate uniformly distributed random numbers over the interval [0,1], and gaussian-distributed numbers, respectively. GAUSS calls RANDU twelve times, thus generating twelve independent uniformly distributed random numbers $x_1, x_2, \ldots, x_{12}$. Then, their sum $x = x_1 + x_2 + \cdots + x_{12}$ will be approximately gaussian, as guaranteed by the central limit theorem. It is interesting to note that the variance of x is unity, as it follows from the fact that the variance of each x_i is 1/12:

$$\sigma_x^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 + \cdots + \sigma_{x_{12}}^2 = \frac{1}{12} + \frac{1}{12} + \cdots + \frac{1}{12} = 1$$

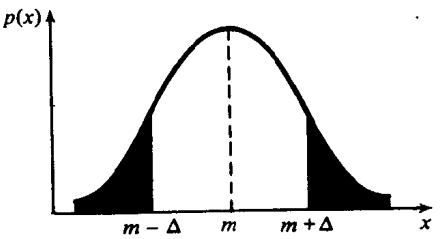
The mean of x is $12/2 = 6$. By shifting and scaling x, one can obtain a gaussian-distributed random number of any desired mean and variance.

1.2 *Chebyshev's Inequality*

The variance σ^2 of a random variable x is a measure of the spread of the x-values about their mean. This intuitive interpretation of the variance is a direct consequence of Chebyshev's inequality, which states that the x-values tend to cluster about their mean in the sense that the probability of a value not occurring in the near vicinity of the mean is small; and it is smaller the smaller the variance. More precisely, for any probability density $p(x)$ and any $\Delta > 0$, the probability that x will fall outside the interval of values $[m - \Delta, m + \Delta]$ is bounded by

$$\text{Prob}[|x - m| \geq \Delta] \leq \frac{\sigma^2}{\Delta^2}$$

(Chebyshev's inequality)



Thus, for fixed Δ, as the variance σ^2 becomes smaller, the x-values tend to cluster more narrowly about the mean. In the extreme limiting case of a deterministic variable $x = m$, the density becomes infinitely narrow, $p(x) = \delta(x - m)$, and has zero variance.

1.3 *Joint and Conditional Densities, and Bayes' Rule*

Next, we discuss random vectors. A pair of two different random variables $\mathbf{x} = (x_1, x_2)$ may be thought of as a vector-valued random variable. Its statistical description is more complicated than that of a single variable and requires knowledge of the joint probability density $p(x_1, x_2)$. The two random variables may or may not have any dependence on each other. It is possible, for example, that if x_2 assumes a particular value, then this fact may influence, or restrict, the possible values that x_1 can then assume.

Example 1.3.1: Suppose that x_1 is related to x_2 by

$$x_1 = 5x_2 + v$$

where v is itself a random variable which is assumed to be independent (see below) of the random variable x_2. Then, the randomness of x_1 arises both from the randomness of x_2 and the randomness of v. But if a value of x_2 has already been realized, then the only randomness left in x_1 will arise only from the random variable v.

A quantity that provides a measure for the degree of dependency of the two variables on each other is the *conditional density* $p(x_1/x_2)$ of x_1 given x_2; and $p(x_2/x_1)$ of x_2 given x_1. These are related by *Bayes' rule*

$$p(x_1, x_2) = p(x_1/x_2)\, p(x_2) = p(x_2/x_1)\, p(x_1)$$

More generally, Bayes' rule for two events A and B is

$$p(A,B) = p(A/B)\, p(B) = p(B/A)\, p(A)$$

The two random variables x_1 and x_2 are *independent* of each other if they do not condition each other in any way; that is, if

$$p(x_1/x_2) = p(x_1) \qquad \text{or} \qquad p(x_2/x_1) = p(x_2)$$

In other words, the occurrence of x_2 does not in any way influence the variable x_1. When two random variables are independent, their joint density factors into the product of single (marginal) densities:

$$p(x_1, x_2) = p(x_1)\, p(x_2)$$

The converse is also true. The correlation between x_1 and x_2 is defined by the expectation value

$$E[x_1x_2] = \int\int x_1x_2\, p(x_1,x_2)\, dx_1dx_2$$

When x_1 and x_2 are independent, the correlation also factors as $E[x_1x_2] = E[x_1]\, E[x_2]$.

The concept of a *random vector* generalizes to any dimension. A vector of N random variables

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

requires knowledge of the joint density

$$p(\mathbf{x}) = p(x_1,x_2, \ldots ,x_N) \tag{1.3.1}$$

for its complete statistical description. The second-order statistics of $\mathbf{x}$ are its mean, its *correlation matrix*, and its *covariance matrix*, defined by

$$\mathbf{m} = E[\mathbf{x}],\ R = E[\mathbf{x}\mathbf{x}^T],\ \Sigma = E[(\mathbf{x} - \mathbf{m})\,(\mathbf{x} - \mathbf{m})^T] \tag{1.3.2}$$

where the superscript T denotes transposition, and the expectation operations are defined in terms of the joint density Eq. (1.3.1); for example,

$$E[\mathbf{x}] = \int \mathbf{x}p(\mathbf{x})d^N\mathbf{x}$$

where $d^N\mathbf{x} = dx_1\, dx_2 \ldots dx_N$ denotes the corresponding N-dimensional volume element. The ijth element of the correlation matrix R is the correlation between the ith random variable x_i with the jth random variable x_j; that is, $R_{ij} = E[x_i\, x_j]$. It is easily shown that the covariance and correlation matrices are related by

$$\Sigma = R - \mathbf{m}\mathbf{m}^T$$

When the mean is zero, R and Σ coincide. Both R and Σ are *symmetric positive semi-definite* matrices.

Example 1.3.2: The probability density of a gaussian random vector $\mathbf{x} = (x_1,x_2, \ldots ,x_N)^T$ is completely specified by its mean $\mathbf{m}$ and covariance matrix Σ; that is,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}\,(\det \Sigma)^{1/2}} \exp\left[-\frac{1}{2}\,(\mathbf{x} - \mathbf{m})^T\, \Sigma^{-1}\, (\mathbf{x} - \mathbf{m}) \right]$$

Example 1.3.3: Under a *linear* transformation, a gaussian random vector remains gaussian. Let $\mathbf{x}$ be a gaussian random vector of dimension N, mean $\mathbf{m}_x$, and covariance Σ_x. Show that the linearly transformed vector

$$\boldsymbol{\xi} = B\mathbf{x} \qquad \text{where } B \text{ is a nonsingular } N \times N \text{ matrix,}$$

is gaussian-distributed with mean and covariance given by

$$\mathbf{m}_\xi = B\mathbf{m}_x, \quad \Sigma_\xi = B\,\Sigma_x B^T \tag{1.3.3}$$

The relationships (1.3.3) are valid also for non-gaussian random vectors. They are easily derived as follows:

$$E[\boldsymbol{\xi}] = E[B\mathbf{x}] = BE[\mathbf{x}]$$

$$E[\boldsymbol{\xi}\boldsymbol{\xi}^T] = E[B\mathbf{x}(B\mathbf{x})^T] = BE[\mathbf{x}\mathbf{x}^T]\,B^T$$

The probability density $p_\xi(\boldsymbol{\xi})$ is related to the probability density $p_x(\mathbf{x})$ by the requirement that, under the above change of variables, they both yield the same elemental probabilities; that is,

$$p_\xi(\xi)\,d^N\boldsymbol{\xi} = p_x(\mathbf{x})\,d^N\mathbf{x} \tag{1.3.4}$$

Since the Jacobian of the transformation from $\mathbf{x}$ to $\boldsymbol{\xi}$ is $d^N\boldsymbol{\xi} = |\det B|d^N\mathbf{x}$, we obtain $p_\xi(\boldsymbol{\xi}) = p_x(\mathbf{x})/|\det B|$. Noting the invariance of the quadratic form

$$(\boldsymbol{\xi} - \mathbf{m}_\xi)^T\,\Sigma_\xi^{-1}\,(\boldsymbol{\xi} - \mathbf{m}_\xi) = (\mathbf{x} - \mathbf{m}_x)^T\,B^T\,(B\,\Sigma_x\,B^T)^{-1}\,B\,(\mathbf{x} - \mathbf{m}_x)$$

$$= (\mathbf{x} - \mathbf{m}_x)^T\,\Sigma_x^{-1}\,(\mathbf{x} - \mathbf{m}_x)$$

and that $\det \Sigma_\xi = \det(B\,\Sigma_x\,B^T) = (\det B)^2 \det \Sigma_x$, we obtain

$$p_\xi(\boldsymbol{\xi}) = \frac{1}{(2\pi)^{N/2}(\det \Sigma_\xi)^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\xi} - \mathbf{m}_\xi)^T\Sigma_\xi^{-1}(\boldsymbol{\xi} - \mathbf{m}_\xi)\right]$$

Example 1.3.4: Consider two zero-mean random vectors $\mathbf{x}$ and $\mathbf{y}$ of dimensions N and M, respectively. Show that if they are *uncorrelated* and *jointly gaussian*, then they are also *independent* of each other. That $\mathbf{x}$ and $\mathbf{y}$ are jointly gaussian means that the $(N + M)$-dimensional joint vector $\mathbf{z} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$ is zero-mean and gaussian; that is,

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{N+M/2}(\det R_{zz})^{1/2}} \exp\left[-\frac{1}{2}\mathbf{z}^T R_{zz}^{-1}\,\mathbf{z}\right]$$

where the correlation (covariance) matrix R_{zz} is

$$R_{zz} = E[\mathbf{z}\mathbf{z}^T] = E\left[\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\overbrace{\mathbf{x}^T,\mathbf{y}^T}\right] = \begin{bmatrix}E[\mathbf{x}\mathbf{x}^T] & E[\mathbf{x}\mathbf{y}^T]\\ E[\mathbf{y}\mathbf{x}^T] & E[\mathbf{y}\mathbf{y}^T]\end{bmatrix} = \begin{bmatrix}R_{xx} & R_{xy}\\ R_{yx} & R_{yy}\end{bmatrix}$$

If $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated; that is, $R_{xy} = E[\mathbf{x}\mathbf{y}^T] = 0$, then the matrix R_{zz} becomes block diagonal and the quadratic form of the joint vector becomes the sum of the individual quadratic forms:

$$\mathbf{z}^T R_{zz}^{-1}\,\mathbf{z} = [\mathbf{x}^T,\mathbf{y}^T]\begin{bmatrix} R_{xx}^{-1} & 0 \\ 0 & R_{yy}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mathbf{x}^T R_{xx}^{-1}\,\mathbf{x} + \mathbf{y}^T R_{yy}^{-1}\,\mathbf{y}$$

Since $R_{xy} = 0$ also implies that $\det R_{zz} = (\det R_{xx})(\det R_{yy})$, it follows that the joint density $p(\mathbf{x},\mathbf{y}) = p(\mathbf{z})$ factors into the marginal densities:

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{z}) = p(\mathbf{x})\,p(\mathbf{y})$$

which shows the independence of $\mathbf{x}$ and $\mathbf{y}$.

Example 1.3.5: Given a random vector $\mathbf{x}$ with mean $\mathbf{m}$ and covariance Σ, show that the best choice of a *deterministic* vector $\hat{\mathbf{x}}$ which minimizes the quantity

$$R_{ee} = E[\mathbf{e}\mathbf{e}^T] = \text{minimum}, \qquad \text{where } \mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$$

is the mean $\mathbf{m}$ itself. That is, $\hat{\mathbf{x}} = \mathbf{m}$. Also show that for this optimal choice of $\hat{\mathbf{x}}$, the actual minimum value of the quantity R_{ee} is the covariance Σ.

This property is easily shown by working with the deviation of $\hat{\mathbf{x}}$ from the mean $\mathbf{m}$; that is, let

$$\hat{\mathbf{x}} = \mathbf{m} + \boldsymbol{\Delta}$$

Then, the quantity R_{ee} becomes

$$\begin{aligned} R_{ee} &= E[\mathbf{e}\mathbf{e}^T] = E[(\mathbf{x} - \mathbf{m} - \boldsymbol{\Delta})(\mathbf{x} - \mathbf{m} - \boldsymbol{\Delta})^T] \\ &= E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T] - \boldsymbol{\Delta}\,E[(\mathbf{x} - \mathbf{m})^T] \\ &\qquad - E[(\mathbf{x} - \mathbf{m})]\,\boldsymbol{\Delta}^T + \boldsymbol{\Delta}\boldsymbol{\Delta}^T \\ &= \Sigma + \boldsymbol{\Delta}\boldsymbol{\Delta}^T \end{aligned}$$

Where we used the fact that $E[(\mathbf{x} - \mathbf{m})] = E[\mathbf{x}] - \mathbf{m} = 0$. Since the matrix $\boldsymbol{\Delta}\boldsymbol{\Delta}^T$ is nonnegative-definite, it follows that R_{ee} will be minimized when $\boldsymbol{\Delta} = 0$, and in this case the minimum value will be Σ.

1.4 *Correlation Canceling*

The concept of correlation canceling plays a central role in the development of many optimum signal processing algorithms, because a correlation canceler is also the *best* linear processor for estimating one signal from another.

Consider two zero-mean random vectors $\mathbf{x}$ and $\mathbf{y}$ of dimensions N and M, respectively. If $\mathbf{x}$ and $\mathbf{y}$ are correlated with each other in the sense that

$R_{xy} = E[\mathbf{x}\mathbf{y}^T] \neq 0$, then we would like to remove such correlations by means of a linear transformation of the form

$$\mathbf{e} = \mathbf{x} - H\mathbf{y} \tag{1.4.1}$$

where the $N \times M$ matrix H must be suitably chosen such that the new pair of vectors $(\mathbf{e},\mathbf{y})$ are no longer correlated with each other; that is, we require

$$R_{ey} = E[\mathbf{e}\mathbf{y}^T] = 0 \tag{1.4.2}$$

Using Eq. (1.4.1), we obtain

$$R_{ey} = E[\mathbf{e}\mathbf{y}^T] = E[(\mathbf{x} - H\mathbf{y})\mathbf{y}^T] = E[\mathbf{x}\mathbf{y}^T] - HE[\mathbf{y}\mathbf{y}^T] = R_{xy} - HR_{yy}$$

Then, the condition $R_{ey} = 0$ immediately implies that

$$H = R_{xy}R_{yy}^{-1} = E[\mathbf{x}\mathbf{y}^T]\,E[\mathbf{y}\mathbf{y}^T]^{-1} \tag{1.4.3}$$

Using $R_{ey} = 0$, the covariance matrix of the resulting vector $\mathbf{e}$ is easily found to be

$$R_{ee} = E[\mathbf{e}\mathbf{e}^T] = E[\mathbf{e}(\mathbf{x}^T - \mathbf{y}^T H^T)] = R_{ex} - R_{ey}H^T = R_{ex} = E[(\mathbf{x} - H\mathbf{y})\mathbf{x}^T]$$

or

$$R_{ee} = R_{xx} - HR_{yx} = R_{xx} - R_{xy}R_{yy}^{-1}R_{yx} \tag{1.4.4}$$

The vector

$$\hat{\mathbf{x}} = H\mathbf{y} = R_{xy}R_{yy}^{-1}\mathbf{y} = E[\mathbf{x}\mathbf{y}^T]\,E[\mathbf{y}\mathbf{y}^T]^{-1}\mathbf{y} \tag{1.4.5}$$

obtained by linearly processing the vector $\mathbf{y}$ by the matrix H is called the *linear regression*, or *orthogonal projection*, of $\mathbf{x}$ on the vector $\mathbf{y}$. In a sense to be made precise later, $\hat{\mathbf{x}}$ also represents the best "copy", or *estimate*, of $\mathbf{x}$ that can be made on the basis of the vector $\mathbf{y}$. Thus, the vector $\mathbf{e} = \mathbf{x} - H\mathbf{y} = \mathbf{x} - \hat{\mathbf{x}}$ may be thought of as the *estimation error*.

Actually, it is better to think of $\hat{\mathbf{x}} = H\mathbf{y}$ not as an estimate of $\mathbf{x}$ but rather as an estimate of *that part* of $\mathbf{x}$ which is correlated with $\mathbf{y}$. Indeed, suppose that $\mathbf{x}$ consists of two parts

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$$

such that $\mathbf{x}_1$ is correlated with $\mathbf{y}$, but $\mathbf{x}_2$ is not; that is, $R_{x_2y} = E[\mathbf{x}_2\mathbf{y}^T] = 0$. Then,

$$R_{xy} = E[\mathbf{x}\mathbf{y}^T] = E[(\mathbf{x}_1 + \mathbf{x}_2)\,\mathbf{y}^T] = R_{x_1y} + R_{x_2y} = R_{x_1y}$$

and therefore,

$$\hat{\mathbf{x}} = R_{xy}R_{yy}^{-1}\mathbf{y} = R_{x_1y}R_{yy}^{-1}\mathbf{y} = \hat{\mathbf{x}}_1$$

The vector $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}} = \mathbf{x}_1 + \mathbf{x}_2 - \hat{\mathbf{x}}_1 = (\mathbf{x}_1 - \hat{\mathbf{x}}_1) + \mathbf{x}_2$ consists of the estimation error $\mathbf{x}_1 - \hat{\mathbf{x}}_1$ of the x_1-part plus the x_2-part. Both of these terms are