02905-16-T

THE UNIVERSITY OF MICHIGAN

COLLEGE OF ENGINEERING
DEPARTMENT OF AEROSPACE ENGINEERING

Technical Report

An Empirical Bayes Technique in Communication Theory

STUART C. SCHWARTZ

Supported by:

National Aeronautics and Space Administration
Grant No. NsG-2-59
Washington, D. C.

Administered through:

October 1966

OFFICE OF RESEARCH ADMINISTRATION · ANN ARBOR

THE UNIVERSITY OF MICHIGAN

COLLEGE OF ENGINEERING
Department of Aerospace Engineering

Technical Report

AN EMPIRICAL BAYES TECHNIQUE IN COMMUNICATION THEORY

Stuart C. Schwartz

ORA Project 02905

supported by

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
GRANT NO. NsG-2-59
WASHINGTON, D. C.

administered through:

OFFICE OF RESEARCH ADMINISTRATION ANN ARBOR

October 1966

TABLE OF CONTENTS

TABLE OF CONTENTS (Concluded)

LIST OF ILLUSTRATIONS

LIST OF APPENDICES

LIST OF PRINCIPAL SYMBOLS

Symbol	Meaning	Defined
A	covariance matrix of dimension k	(3.1.3)
a_j	Fourier coefficients of $f(x)$ in the L_2 expansion	(2.4.2)
$\hat{a}_{jn}$	estimate of a_j at the n-th stage	(2.4.4)
$\alpha(z)$	distribution function of the random variable Z	
$\alpha_{m-\ell}(z_1,z_2)$	second-order (stationary) distribution of Z_1 and Z_2	above (2.2.28)
B_τ	cross-covariance matrix of dimension k	(3.1.3)
d_j	coefficients associated with $\alpha(z)$ in the eigenfunction representation	(2.5.10)
$\tilde{D}_{m-\ell}$	averages involving the differences of the bivariate distribution and univariate distribution	(2.2.10)
$D_{m-\ell}$		(2.2.29)
$\tilde{D}_{\tau,2}$		(2.2.35)
$D_{\tau,2}$		(2.2.41)
$f(x)$	overall density function of the observation, $f(x) = \int g(x-z;\sigma)d\alpha(z)$	(2.1.4)
$f_i(x)$	density function of the observation under the i-th hypothesis	
$f_q(x)$	series representation for $f(x)$ truncated at q+1 terms	(2.4.24) or (2.5.25)
$\hat{f}_n(x)$	estimate of $f(x)$ after the n-th observation	
$F_n(x)$	empirical distribution function	(2.2.3)

Symbol	Meaning	Defined
$g(x-y;\sigma)$	gaussian density function with mean value y and standard deviation σ. Designated also as $g(\frac{x-y}{\sigma})$ when dealing with the Hermite polynomials	
$g_2(n_1,n_2;\sigma,\rho)$	bivariate gaussian density with correlation coefficient ρ and both variables having the same standard deviation σ	
$H_j(x)$	Hermite polynomial associated with the weight $g^2(x)$	(A.11)
$He_j(x)$	Hermite polynomial associated with the weight $g(x)$	(A.3)
J_n or $J_{n'}$	mean integrated square error	(2.3.30) or (2.5.20)
$K(y)$	kernel function	(2.3.4)
ℓ,m	used as subscripts to denote the ℓ-th and m-th interval, observation, etc...	
M_T	covariance matrix of dimension 2k	(3.1.3)
$M_{(\ )}(v)$	Fourier transform of the quantity appearing in the subscript	
n	used as a subscript to denote the present observation, decision	
N	covariance matrix of dimension 2k	(3.2.2)
N_i	i-th gaussian noise sample	
O	big O notation, e.g., $V(f_n(x)) = O(1/n)$ means $V(f_n(x)) \leq a/n$ where a is a positive constant	
P_e	probability of error of the one-stage procedure	

LIST OF PRINCIPAL SYMBOLS (Concluded)

Symbol	Meaning	Defined
P_{en}	probability of error of the empirical procedure	
$\varphi_j(x)$	orthonormal Hermite functions	(2.4.3)
$q(n)$	positive integer depending on n	
$\rho_{\ell-m}$	correlation coefficient of the ℓ and m gaussian noise samples	
ρ_*	$\max_{\tau \geq 1} \lvert\rho_\tau\rvert$	
$R(\tau)$	autocorrelation function of the gaussian process	
$s(x)$	a particular exponential function	(2.5.6)
σ	standard deviation of the noise samples, $\sigma^2=R(0)$	
σ_1	an arbitrary constant, which in section 2.5, is chosen greater than σ	
$T(x)$	test function	
$t(x)$	decision function	
τ	an index to denote the difference $\tau=m-\ell$	
$V(\ )$	variance of the indicated quantity	
X_ℓ	denotes the observation $X_\ell=N_\ell+Z_\ell$	
ξ^2	$\xi^2 = (\sigma_1^2-\sigma^2)/(\sigma_1^2+\sigma^2)$	(2.5.3)
$\wedge$	designates the estimate of the quantity appearing under it	

LIST OF CONSTANTS

The following symbols are used for constants or bounds.

Symbol	Defined
B_1	below equation (2.2.20)
B_2	equation (2.2.24) or (2.2.25)
B_3	equation (2.2.44)
B_4	equation (2.3.10)
B_5	below equation (2.3.16)
B_6	below equation (2.4.19)
B_7	above equation (2.3.21)
$B_8(k)$	equation (3.2.41)
b_1	below equation (2.3.25)
b_2	below equation (2.3.25)
b_3	above equation (2.3.26)
c_1	above equation (A.30)
c_2	equation (2.4.7)
c_3	equation (2.4.10)
c_4	equation (2.5.16)
c_5	equation (2.5.17)
c_6	equation (2.5.21)
c_7	equation (2.5.33)
$c_8(i)$	equation (2.6.8)

ABSTRACT

This thesis is concerned with an empirical Bayes procedure and its application to communication theory. The communication problem is one in which a sequence of information bearing signals is either assumed to be a stationary random process or distorted by a stationary random process. In either case, the underlying probability structure is unknown. The message sequence is then added to correlated gaussian noise. The statistical inference problem is to extract information from each member of the observation sequence, i.e., make a decision as to the presence of a particular signal. The empirical Bayes procedure utilizes all past observations to obtain consistent estimates of the unknown distributions or related quantities. These estimates are then used to form a sequence of test functions which is evaluated using only the present observation. It is shown that the sequence of test functions converges to the test function one would use if all distributions were known and if the observations were independent. For a minimum probability of error criterion, the resulting difference in error probabilities is dominated by a quantity proportional to the mean-square error in the estimate of the test function.

In particular, we consider the class of problems where the marginal density function of an observation is the convolution of a gaussian density function and an unknown distribution, $f(x) = \int g(x-z;\sigma)d\alpha(z)$. By suitably interpreting $\alpha(z)$, a variety of communication problems are included. Much of this study is concerned with obtaining consistent estimates of $f(x)$ given the sequence of dependent, identically distributed random variables $X_i=N_i+Z_i$, $i=1,\ldots n$. Three techniques are presented: a kernel method which is similar to the procedure used for estimating a spectral density, an orthogonal expansion for $f(x)$ in Hermite functions, and an eigenfunction representation obtained by solving an eigenfunction problem associated with the integral equation for $f(x)$. For all three methods, we calculate the bounds on the mean-square error in the estimate of $f(x)$. A typical result is: if the autocorrelation function of the gaussian noise is absolutely integrable and eventually monotonically decreasing, and if the sequence Z_i is M-dependent, the rate of convergence of the estimates is the same as in the case of independent observations. The rate is $O(1/n^{4/5})$ for the kernel method. For the orthogonal expansion, with the r-th absolute moment of Z finite, the rate is $O(1/n^{(r-2)/r})$. With the eigenfunction representation, we estimate a quantity related to $f(x)$ and obtain the rate $O(\ln^2 n/n)$. The techniques are then extended to the case of estimating a k-variate density function $f(x_1 \ldots x_k)$.

These results allow us to bound the rate of convergence of the risk incurred using the empirical procedure in a number of communication problems. The problems considered are: communication through an unknown, stationary, random channel when learning samples (channel sounding signals) are available, communication through an unknown random multiplicative channel, and the transmission of known signals with unknown a priori probabilities.

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

This thesis is concerned with a class of hypothesis testing problems in which not all pertinent statistics are known, but where the observer is repeatedly faced with the same decision problem. The type of problem we want to discuss is one in which a sequence of information bearing signals is assumed to be, or distorted by, a stationary random process whose underlying probability structure is unknown. The message sequence is then added to correlated gaussian noise. The statistical inference problem is to extract information from each member of the observation sequence, i.e., make a decision as to the presence of a particular signal. The empirical Bayes technique which we shall discuss involves the use of accumulated past observations to obtain consistent estimates of the unknown distributions or related quantities. These estimates are then used to form a sequence of test functions which converges to the test function one would use if all pertinent distributions were known and if the sequence of observations were independent. These remarks are perhaps best clarified by a simple example.

Suppose we have an observation $X=N+Z$, where N is a gaussian random variable with mean zero and standard deviation equal to one. Z is assumed to a random variable which takes on the values 0 and 1 with probability p_0 and $p_1=1-p_0$, respectively. We take Z independent of N.

Designate the distribution of Z by $\alpha(z)$ and let the gaussian density with a standard deviation equal to 1 be denoted by $g(x;1)$ The density function of the observation X is then written.

$$\begin{aligned} f(x) &= \int g(x-z;1)\, d\alpha(z) \\ &= p_0 g(x;1) + p_1 g(x-1;1) \end{aligned} \tag{1.1.1}$$

We want to test whether Z=0 or 1 with a minimum probability of error criterion. The optimum test procedure is known to be a likelihood ratio test with a threshold of one. Using the logarithm of the likelihood ratio, an equivalent procedure is to evaluate the function.

$$\begin{aligned} T(x) &= x-1/2 + 2\ln\left(\frac{1-p_1}{p_1}\right) \\ &= x-c, \end{aligned} \tag{1.1.2}$$

and compare it to a zero threshold. The test procedure which minimizes the probability of error is to choose H_1 (Z=1) if $T(x) \geq 0$ and H_0 if $T(x) < 0$. Let $G(x)$ denote the cumulative gaussian distribution function. Then, the probability of an incorrect decision is given by

$$P_e = p_1\, G(c-1) + (1-p_1)(1-G(c)). \tag{1.1.3}$$

P_e as a function of p_1 is called the Bayes envelope function and is the minimum probability of error attainable. A plot of this function is given in Figure 1.

Suppose p_1 is unknown and that we have a "good" estimate which we denote by $\hat{p}_1$. Then, we might use the test function

$$\hat{T}(x) = x - 1/2 + 2\ln\left(\frac{1-\hat{p}_1}{\hat{p}_1}\right) \qquad (1.1.4)$$

$$= x - \hat{c}$$

and compare this quantity to a zero threshold. The reason for this is that if $\hat{p}_1$ is close to p_1, $\hat{T}(x)$ ought to be close to the Bayes test function $T(x)$. This is in fact the case as can be seen by calculating the probability of error as a result of using $\hat{T}(x)$. Defining this probability of error as $P(\hat{p}_1, p_1)$, a straightforward calculation yields

$$P(\hat{p}_1, p_1) = p_1 G(\hat{c}-1) + (1-p_1)(1-G(\hat{c})). \qquad (1.1.5)$$

A plot of $P(\hat{p}_1, p_1)$ versus p_1 for different values of $\hat{p}_1$ is also given in Figure 1.

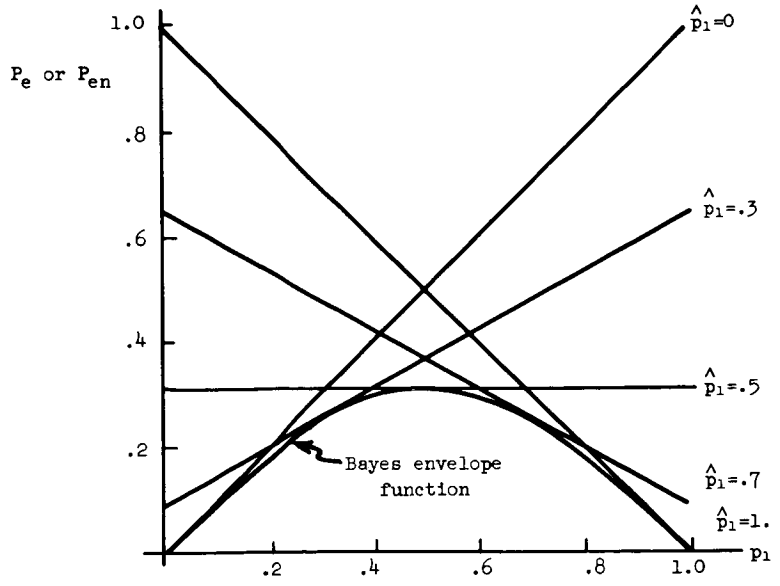

Figure 1. Probability of error vs. p_1.

Now assume we are repeatedly faced with the same decision problem; we observe the sequence of stationary random variables $X_\ell, \ell=1,2,\ldots n$, and for each observation X_ℓ we are to decide whether H_0 or H_1 is the true state of nature. Prior to making a decision based on the observation X_n, we would use X_n to update the estimate of p_1. For this example, a convenient estimate of p_1 is

$$\hat{p}_{1n} = \frac{1}{n} \sum_{\ell=1}^{n} X_\ell \tag{1.1.6}$$

Since $\hat{p}_{1n}$ is a function of the observations, $P(\hat{p}_{1n}, p_1)$ is a random variable. We define the average value of $P(\hat{p}_{1n}, p_1)$ by

$$P_{en} = E\{P(\hat{p}_{1n}, p_1)\}. \tag{1.1.7}$$

For the above procedure to be useful, we should have

$$\lim_{n\to\infty} P_{en} = P_e. \tag{1.1.8}$$

That is, on the average, as the number of observations (and decisions) increases, the probability of error P_{en} should approach P_e. An estimate of how close P_{en} is to P_e as a function of n would also be of value.

In the next section we will show that for (8) to hold we need $\hat{p}_{1n}$ converging in probability to p_1.[1] We also show that if $\hat{p}_{1n}$ converges in mean-square to p_1, a knowledge of the mean-square error provides a bound on the difference $P_{en}-P_e$.

---

[1] We will refer to equations within the section by the last index. In referring to an equation in another section we will use all three indices.

Let us calculate the mean-square error for the estimate $\hat{p}_{1n}$. Using (1), we see that the estimate is unbiased

$$E\{\hat{p}_{1n}\} = p_1 \tag{1.1.9}$$

Assume the sequence of observations $\{X_\ell\}$ is independent. Then, the mean-square error is

$$E\,(\hat{p}_{1n}-p_1)^2 = V(\hat{p}_{1n}) = \frac{1+p_1(1-p_1)}{n} \tag{1.1.10}$$

The estimate $\hat{P}_{1n}$ converges in mean-square at the rate $O(1/n)$.[1]

If, on the otherhand, we assume that the gaussian noise samples are correlated, $E(N_\ell N_m) = \rho_{m-\ell}$, we have

$$V(\hat{p}_{1n}) = \frac{1+p_1(1-p_1)}{n} + \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}\rho_{m-\ell} \tag{1.1.11}$$

From the stationarity assumption, we can write

$$V(\hat{P}_{1n}) = \frac{1+p_1(1-p_1)}{n} + \frac{2}{n}\sum_{\tau=1}^{n}(1-\frac{\tau}{n})\rho_\tau \tag{1.1.12}$$

Assuming the correlation coefficients ρ_τ are absolutely summable,

$$\sum_{\tau=1}^{\infty}|\rho_\tau| \leq B_2 < \infty, \tag{1.1.13}$$

we have

---

[1] We use the standard big O notation to write $V(\hat{p}_{1n})=O(1/n)$, which is taken to mean that there is a constant a for which $V(\hat{p}_{1n}) \leq a/n$.

$$V(\hat{p}_{1n}) \leq \frac{1+p_1(1-p_1)}{n} + \frac{2}{n}\sum_{\tau=1}^{n} |\rho_\tau| \leq \frac{1}{n}\,(1+p_1(1-p_1)+2B_2). \tag{1.1.14}$$

Hence, if (13) holds, $\hat{p}_{1n}$ converges in mean-square also at the rate $O(1/n)$.

Now, for the case of independent samples, $T(x)$ as given by (2) is the optimum test function; the test procedure using (2) results in the minimum probability of error, P_e. Then, using $\hat{p}_{1n}$ in (4) we have

$$\hat{T}_n(x) = x - 1/2 + 2\ln\left(\frac{1-\hat{p}_{1n}}{\hat{p}_{1n}}\right). \tag{1.1.15}$$

If, as a result of using $\hat{T}_n(x)$, P_{en} converges to P_e, we say the sequence of test functions is asymptotically optimal. This definition was introduced by Robbins [30].[1]

When the sequence of observations is dependent we will still use the test function given by (15). With $V(\hat{p}_{1n})=O(1/n)$, we would expect that P_{en} converges to P_e at the same rate as for the case of independent observations. Now P_e is no longer the minimum probability of error attainable since we do not base the present decision on all past observations. We do, however, use all past observations to form an estimate of $T(x)$. For this case, we shall call $T(x)$, as given by (2), the "optimum one-stage" test function. We will be concerned with the convergence of $\hat{T}_n(x)$ to this one-stage test function. Clearly, the one-stage

[1]Numbers in square brackets refer to references listed at the end of the report.

test can be modified by basing a decision on a specified number of observations. Then, at the expense of increasing the number of hypotheses to be tested, the probability of error would tend to decrease. In general, this procedure is still suboptimum. We do not discuss it further.

We have called the learning and test procedure an empirical Bayes procedure because the decision problem is one of making an inference concerning the presence of a random variable (or process) Z which is distributed according to some a priori distribution $\alpha(z)$ and which we will take as unknown. With dependent observations, which is the case we will study, the procedure is neither an optimum (Bayes) nor asymptotically optimum procedure.

We now generalize the above problem and establish bounds on the convergence of P_{en} to P_e.

1.2 THE EMPIRICAL BAYES PROCEDURE

We let the parameter λ represent the hypothesis in effect when the random variable X is observed.[1] λ takes on the values "0" and "1" with probability p_o and $p_1=1-p_o$, respectively. The observed random variable X is governed by the density function $f_i(x)$ when $\lambda=i$, $i=0,1$. The density function $f_i(x)$ will, in general, be the convolution of a gaussian density and some distribution function. The marginal, or overall density function of the observation is

$$f(x) = p_o f_o(x) + p_1 f_1(x) \qquad (1.2.1)$$

[1] In this formulation, and in the proof of asymptotic optimality for independent observations, we follow Robbins[30].

The choice of deciding between the two hypotheses is made by a decision function $t(x)$; $t(x)$ is defined on the space of observations and takes the values 0 or 1, according to which hypothesis we believe is active. A loss function $L(t(x),\lambda)$ is also defined as the loss incurred when we make the decision $t(x)$ and λ is the true parameter. We take the loss of a correct decision to be zero, $L(0,0)=L(1,1)=0$, and define

$$L(1,0) = a_0 > 0, L(0,1) = a_1 > 0.^{1} \qquad (1.2.2)$$

Letting $b(\lambda) = L(0,\lambda) - L(1,\lambda)$, we can write the loss incurred by using $t(x)$ as

$$L(t(x),\lambda) = L(0,\lambda) - t(x)b(\lambda). \qquad (1.2.3)$$

For any decision $t(x)$, the expected loss as a function of the parameter λ is

$$R(t,\lambda) = \int_X L(t(x),\lambda) f_\lambda(x) dx. \qquad (1.2.4)$$

The expected or overall risk is defined as

$$R(t,\pi) = \int_\lambda R(t,\lambda)\, d\pi(\lambda)$$

$$= \int_\lambda \int_X L(t(x),\lambda)\, f_\lambda(x)\, d\pi(\lambda)$$

---

[1] We will carry a_0 and a_1 along in the development even though we are interested in a minimum probability of error criterion for which $a_0 = a_1 = 1$.

where $\pi(\lambda)$ is the distribution forλ. We can write the expected risk in the form

$$R(t,\pi) = p_1a_1 - \int_x t(x)\,[p_1a_1f_1(x) - p_oa_of_o(x)]dx. \quad (1.2.5)$$

If we denote the test function $T(x)$ by

$$T(x) = p_1a_1f_1(x) - p_oa_of_o(x), \quad (1.2.6)$$

then (5) becomes

$$R(t,\pi) = p_1a_1 - \int_x t(x)T(x)\,dx, \quad (1.2.7)$$

and the procedure to minimize the overall risk is to choose

$$\begin{aligned} t_B(x) &= 1 \quad \text{if} \quad T(x) \geq 0 \\ &= 0 \quad \text{if} \quad T(x) < 0. \end{aligned} \quad (1.2.8)$$

The decision function defined in this manner is the Bayes decision function (with respect to the distribution π), and the Bayes (minimun) risk is

$$R(t_B,\pi) = p_1a_1 - \int T(x)^+ \,dx, \quad (1.2.9)$$

where $T(x)^+ = T(x)$ if $T(x) \geq 0$ and 0 if $T(x) < 0$.

Now suppose that the test function $T(x)$ is unknown and that we are repeatedly faced with the same decision problem. (Both p_i and $f_i(x)$ may be taken as unknown.) At the n-th decision, we have observed the sequence of stationary random variables $X_\ell, \ell=1,2,\ldots n$, and we want to

decide whether $\lambda_n=0$ or $\lambda_n=1$. The values of λ_ℓ (the states of nature), $\ell=1,2,\ldots n-1$, may or may not be known, i.e., we may not know whether our previous decisions were correct. We agree to use only the present observations X_n to make a decision as to the value of λ_n, but we use all past observations to determine a decision function t_n which takes on the values 0 or 1. The decision function is now defined on the space of all past observations. We designate this decision function by $t_n(x_1,x_2,\ldots x_n)$ or, for notational convenience, by $t_n(x_n)$.

Let $E_{n\lambda n}$ denote the mathematical expectation with respect to all the random variables $X_1\ldots X_n\lambda_n$, and $E_{n|\lambda n}$ denote the conditional expectation given λ_n. With the loss given by (3), the expected loss at the n-th stage is

$$R(t_n,\lambda_n) = E_{n|\lambda n}(L(t_n(X_n),\lambda_n)) \tag{1.2.10}$$

and the overall loss is given by

$$\begin{aligned} R(t_n,\pi) &= \int R(t_n,\lambda)\,d\pi(\lambda) \\ &= E_{n\lambda n}(L(t_n(x_n),\lambda_n)). \end{aligned} \tag{1.2.11}$$

This can also be written as

$$R(t_n,\pi) = p_1a_1 - E_{n\lambda n}(t_n(X_n)b(\lambda_n)) \tag{1.2.12}$$

Let $E_{x\lambda}$ denote the expectation with respect to the pair of random variables (x,λ). If $\lim_{n\to\infty} E_{n\lambda n}\{(t_n(X_n)b(\lambda_n))\}=E_{x\lambda}\{(t_B(X)b(\lambda))\}$ then, in view of (4)-(6), we have

$$\lim_{n\to\infty} R(t_n,\pi) = R(t_B\pi). \qquad (1.2.13)$$

A sequence of decision functions $\{t_n(x_1\ldots x_n)\}$ such that (13) is satisfied is said to be asymptotically optimal. This is the definition Robbins adopted for the case of independent observations. We shall now obtain a bound on the convergence of (13) and also investigate the case where the observations are dependent.

1.2a Independent Observations

Consider the second expression in (12):

$$E_n\lambda_n(t_n(X_n)b(\lambda_n)) = p_oE_n|_{\lambda=o}(t_n(X_n)b(0)) + p_1E_n|_{\lambda=1}(t_n(X_n)b(1)).$$

In view of the independence and stationarity assumption, and the definition of $b(\lambda)$, from (1) we have

$$p_oE_n|_{\lambda_n=o}(t_n(X_n)b(0)) = -a_op_o\int f_o(x_n)$$

$$\left\{\overbrace{\int\ldots\int}^{n-1} t(x_1\ldots x_n)f(x_{11})\ldots f(x_{n-1})dx_1\ldots dx_{n-1}\right\}dx_n$$

and

$$p_1E_n|_{\lambda_n=1}(t_n(X_n)b(1)) = p_1a_1\int f_1(x_n)$$

$$\left\{\overbrace{\int\ldots\int}^{n-1} t_n(x_1\ldots x_n)f(x_{n-1})dx_1\ldots dx_{n-1}\right\}dx_n$$

Using the definition for $T(x)$, we can write

$$E_n\lambda_n(t_n(X_n)b(\lambda_n)) = \int T(x_n)$$

$$\left\{\int\ldots\int t_n(x_1\ldots x_n)f(x_1)\ldots f(x_{n-1})dx_1\ldots dx_{n-1}\right\}dx_n$$

Define a sequence of test functions

$$T_n(X_n) = T_n(X_1, X_2, \ldots, X_{n-1}, X_n; X_n).^1$$

$T_n(x_n)$ is a function of x_n, whose functional form depends on the variables (observations) $x_1 \ldots x_n$. Suppose that for almost every fixed x and arbitrary ϵ,

$$\lim_{n\to\infty} \Pr\{|T_n(X_1, \ldots X_{n-1}, x; x) - T(x)| < \epsilon\} = 1, \quad (1.2.14)$$

i.e., $T_n(x_n)$ converges in probability to $T(x)$ for almost every fixed x. Further, define the sequence of decision functions by

$$t_n(x_1 \ldots x_n) = 1 \text{ if } T_n(x_1 \ldots x_n; x_n) \geq 0$$
$$= 0 \text{ otherwise} \quad (1.2.15)$$

We then have

$$E_{n\lambda n}(t_n(X_n) b(\lambda_n)) = \int T(x) [\Pr\{T_n(X_1 \ldots X_{n-1}, x; x) \geq 0\}] \, dx,$$

and since $\int |T(x)| dx < \infty$, it follows from the dominated convergence theorem and (14) that

$$\lim_{n\to\infty} E_{n\lambda n}(t_n(X_n) b(\lambda_n)) = \int T(x) \lim_{n\to\infty} \Pr\{T_n(X_1 \ldots X_{n-1}, x; x) \geq 0\} \, dx$$
$$= \int T(x)^+ dx.$$

---

[1] In this section we drop the $\wedge$ notation on $\hat{T}_n(X_n)$.

Theorem 1.2.1 (Robbins [30], p. 201): With $T_n(x_n)$ such that (14) is true and with $t_n(x_n)$ defined by (15), the sequence of test functions is asymptotically optimal in the sense that

$$\lim_{n\to\infty} R(t_n,\pi) = R(t_B,\pi) \ .$$

Perhaps the most convenient way to obtain a rate of convergence is to assume that the sequence of test functions converges in mean-square to $T(x)$, uniformly in x. Suppose that for almost every x, the inequality

$$E_{n-1}\{|T_n(X_1\ldots X_{n-1},x;x)-T(x)|\}^2 \leq \beta_n^2 \qquad (1.2.16)$$

is satisfied and that $\lim_{n\to\infty} \beta_n=0$.[1] Then, from the Chebyshev inequality, we have convergence in probability with the bound

$$Pr\{|T_n(X_1\ldots X_{n-1},x;x)-T(x)| \geq \epsilon \} \leq \beta_n^2 /\epsilon^2$$

for a.e.x.

Notice that by definition $R(t_n,\pi) \geq R(t_B,\pi)$. From (9) and (12), the difference in risks is given by

$$0 \leq R(t_n,\pi)-R(t_B,\pi) = \int T(x)^+dx - \int T(x)[Pr\{T_n(X_1,X_2\ldots X_{n-1},x;x) \geq 0\}]dx$$

Define $T(x)^- = T(x)$ if $T(x) < 0$ and $T(x)^- = 0$ if $T(x) \geq 0$. We then have

---

[1] E_{n-1} denotes the expectation with respect to the first (n-1) random variables, $X_1\ldots X_{n-1}$.

$$0 \leq R(t_n,\pi)-R(t_B,\pi) = \int T(x)^+dx - \int T(x)^+[\Pr\{T_n(x) \geq 0\}]dx$$

$$- \int T(x)^-[\Pr\{T_n \geq 0\}]dx \qquad (1.2.17)$$

$$= \int T(x)^+[\Pr\{T_n(x) < 0\}]dx - \int T(x)^-[\Pr\{T_n(x) \geq 0\}]dx.$$

Let $A = \{x: 0 \geq T(x) \geq \epsilon\}$ for arbitrary positive ϵ and consider the first expression on the right side of (17).

$$\int T(x)^+[\Pr\{T_n(x) < 0\}]dx = \int_A T(x)^+[\Pr\{T_n(x) < 0\}]dx + \int_{A_c} T(x)^+[\Pr\{T_n(x) < 0\}]dx$$

For x contained in A, it follows from the bound below (16) that $\Pr\{T_n \leq 0\} = \Pr\{T_n - T \leq -T\} \leq \Pr\{T_n - T \leq -\epsilon\} \leq \beta_n^2/\epsilon^2$. Hence, the first integral in the above expression is bounded by $\frac{\beta_n^2}{\epsilon^2}\int_A T(x)^+dx$. For the second integral, assuming $a_1 > a_0$, we have

$$\int_{A_c} T(x)^+[\Pr\{T_n(x) < 0\}]dx \leq \int_{A_c} T(x)^+dx$$

$$\leq a_1 \int_{A_c}[p_1 f_1(x) + p_0 f_0(x)]dx = a_1 \Pr[0 \leq T(x) \leq \epsilon\}$$

$$= a_1 \delta_1(\epsilon) .$$

Collecting results, we have

$$\int T(x)^+[\Pr\{T_n(x) < 0\}]dx \leq \frac{\beta_n^2}{\epsilon^2}\int T(x)^+dx + a_1\delta_1(\epsilon) .$$

In a similar manner, the second integral on the right side of (17) is bounded by

$$- \int T(x)^-[\Pr\{T_n(x) \geq 0\}]dx \leq - \frac{\beta_n^2}{\epsilon^2}\int T(x)^-dx + a_1\delta_2(\epsilon)$$

where $\delta_2(\epsilon) = Pr\{ -\epsilon \leq T(x) \leq 0\}$. We now have,

Corollary 1.2.1: Assume that the sequence of test functions T_n converges in mean-square as given by (16). Then, the risk at the n-th decision is bounded by

$$0 \leq R(t_B,\pi)-R(t_n,\pi) \leq \frac{\beta_n^2}{\epsilon^2} \int|T(x)|dx + \delta(\epsilon) ,$$

where $\delta(\epsilon) = Pr\{|T(x)| < \epsilon\}$.

We shall now derive the same bound for the case of dependent observations and also give conditions on $f_o(x)$ and $f_1(x)$ so that $\delta(\epsilon)$ can be made arbitrarily small.

1.2b Dependent Observations

Let: $f(x_1,\ldots,x_n)$ denote the(n-variate) density function of the random variables $X_\ell, \ell=1,\ldots n$; $f(x_1,\ldots,x_{n-1}|x_n,\lambda_n)$ be the conditional density function of the first n-1 random variables given λ_n and x_n; and $f(x_1,\ldots,x_n|\lambda_n)$ designate the density function of the variables $x_1\ldots,x_n$, given λ_n. In analogy to the previous development, we have

$$R_n(tn,\pi) = E_{n\lambda_n}(L(t_n(X_n),\lambda_n)) = p_1a_1 - E_{n\lambda_n}(t_n(X_n)b(\lambda_n)), \quad (1.2.18)$$

and

$$E(t_n(X_n)b(\lambda_n)) = p_1a_1E_{n|\lambda_{n=1}}(t_n(X_n)) - p_oa_oF_{n|\lambda_{n=o}}(t_n(X_n)).$$

We write the first expectation on the right side as

$$E_{n|\lambda_{n=1}}(t_n(X_n)) = \int \cdots \int t_n(x_1 \cdots x_n) f(x_1 \cdots x_n | \lambda_n = 1)\, dx_1 \cdots dx_n$$

$$= \int f_1(x_n) \left\{ \int^{n-1} \cdots \int t_n(x_1 \cdots x_n) f(x_1 \cdots x_{n-1} | x_n, \lambda_n = 1)\, dx_1 \cdots dx_{n-1} \right\} dx_n$$

The conditional density $f(x_n|\lambda_n=1)$ is written as $f_1(x_n)$. The second expectation, $E_{n|\lambda_n=0}$, is written with $f_0(x_n)$ in place of $f_1(x_n)$ and $f(x_1,\ldots,x_{n-1}|x_n,\lambda_n=0)$ in place of $f(x_1,\ldots x_{n-1}|x_n,\lambda_n=1)$.

Using these expressions in (18), the empirical risk becomes

$$R(t_n,\pi) = p_1 a_1 - \left\{ \int p_1 a_1 f_1(x_n) \left[\int^{n-1} \cdots \int t_n(x_1 \cdots x_n) \right.\right.$$

$$\left. f(x_1 \cdots x_{n-1} | x_n, \lambda_{n=1})\, dx_1 \cdots dx_{n-1} \right] dx_n$$

$$- \int p_0 a_0 f_0(x_n) \left[\int^{n-1} \cdots \int t_n(x_1 \cdots x_n) f(x_1 \cdots x_{n-1} | x_n, \lambda_n = 0)\, dx_1 \cdots \right.$$

$$\left.\left. dx_{n-1} \right] dx_n \right\},$$

and upon using the definition of $T(x)$, the risk is expressed as

$$R(t_n,\pi) = p_1 a_1 - \left\{ \int T(x) \left[\int^{n-1} \cdots \int t_n(x_1 \cdots x_{n-1}, x_n) \right.\right.$$

$$\left. f(x_1 \cdots x_{n-1} | x_n = x, \lambda_n = 1)\, dx_1 \cdots dx_{n-1} \right] dx$$

$$- \int p_0 a_0 f_0(x) \left[\int^{n-1} \cdots \int t_n(x_1 \cdots x_{n-1}, x) \left\{ f(x_1 \cdots x_{n-1} | x_n = x, \lambda_n = 0) - \right.\right.$$

$$\left.\left. f(x_1 \cdots x_{n-1} | x_n = x, \lambda_n = 1) \right\} dx_1 \cdots dx_{n-1} \right] dx \Bigg\}. \qquad (1.2.19)$$

To demonstrate convergence, in contrast to (14), we now need to require convergence in probability conditioned on the n-th pair of random variables x_n and λ_n,

$$\lim_{n\to\infty} \Pr\left\{ |T_n(x_1 \ldots x_{n-1}, x; x) - T(x)| < \epsilon \,|\, X_n = x, \lambda_n = i \right\} =$$

$$0, \quad i = 0,1. \tag{1.2.20}$$

Clearly, this is satisfied if we have (conditional) mean-square convergence for a.e.x,

$$\lim_{n\to\infty} E_{n-1|\lambda_n=i}\left(|T_n - T|^2\right) = \lim_{n\to\infty} \int^{n-1} \cdots \int (T_n(x_1 \ldots x_{n-1}, x; x) - T(x))^2$$

$$f(x_1 \ldots x_{n-1} | X_n = x, \lambda_n = i)\, dx_1 \ldots dx_{n-1} =$$

$$0, \quad i = 0,1. \tag{1.2.21}$$

This condition, however, is difficult to verify. Under the assumption that the marginal density functions $p_i f_i(x)$, $i=0,1$, do not equal zero for almost all x, (21) is implied by the inequality

$$E_n(T_n - T)^2 = \int \cdots \int (T_n(x_1 \ldots x_{n-1}, x_n; x_n) - T(x_n))^2 f(x_1 \ldots x_n)\, dx_1 \ldots$$

$$dx_n \leq \gamma_n^2 \tag{1.2.22}$$

where $\lim_{n\to\infty} \gamma_n = 0$. This average is considerably easier to obtain.

Then, it is easy to see that the empirical risk converges to the risk incurred by using the one-state test:

$$\lim_{n\to\infty} R(t_n,\pi) = p_1a_1 - \int T(x) \lim_{n\to\infty} \Pr\Big\{ T_n(X_1 \ldots X_{n-1}, x; x) \geq 0 \,\Big|\, \lambda_n = 0,\ X_n = x \Big\} dx = p_1a_1 - \int T(x)^+ dx = R(t_B,\pi).$$

To get a rate of convergence, we proceed differently than in the independent case; the reason being that the bound in (22) does not imply a useful bound for (21) and (20).

From (5), we have

$$R(t_B,\pi) = p_1a_1 - E_{x\lambda}(t_B(X)b(\lambda)),$$

and since $R(t_B,\pi)$ depends only on the present pair of random variables (x_n,λ_n), we can write the difference of (18) and $R(t_B,\pi)$ as

$$0 \leq R(t_n,\pi) - R(t_B,\pi) =$$

$$E_{n\lambda_n}(b(\lambda_n)(t_B(X_n) - t_n(X_1 \ldots X_n))). \qquad (1.2.24)$$

Since $b(\lambda)$ is a bounded function and assuming $a_1 > a_0$, (24) is dominated by

$$0 \leq R(t_n,\pi) - R(t_B,\pi) \leq a_1 E_n |t_B(X_n) - t_n(X_1 \ldots X_n)|. \qquad (1.2.25)$$

The functions t_B and t_n take on the values 0 or 1. Hence, the contributions to the expectation are the two cases where $t_B \neq t_n$. We have, from (8) and (15),

$$0 \leq R(t_n,\pi) - R(t_B,\pi) \leq a_1 \Pr\Big\{ T_n(X_1, X_2, \ldots X_n; X_n) \geq 0,\ T(X_n) < 0 \Big\}$$

$$+ a_1 \Pr\Big\{ T_n(X_1, X_2 \ldots X_n; X_n) < 0, T(X_n) \geq 0 \Big\}.$$

Let $\epsilon > 0$ be an arbitrary constant and consider the first expression on the right side:

$$\Pr\left\{T_n \geq 0,\ T < 0\right\} = \Pr\left\{T_n \geq 0,\ T \leq -\epsilon\right\} + \Pr\left\{T_n \geq 0,\ -\epsilon < T < 0\right\}.$$

Assuming (22) holds, we have

$$\Pr\left\{T_n \geq 0,\ T \leq -\epsilon\right\} \leq \Pr\left\{|T_n - T| \geq \epsilon\right\} \leq \gamma_n^2/\epsilon^2.$$

Letting $\delta_1(\epsilon) = \Pr\left\{-\epsilon < T(x_n) < 0\right\}$, it follows that

$$\Pr\left\{T_n \geq 0,\ -\epsilon < T < 0\right\} = \delta_1(\epsilon)\ \Pr\left\{T_n \geq 0 \mid -\epsilon < T < 0\right\} \leq \delta_1(\epsilon).$$

Hence, we have

$$\Pr\left\{T_n \geq 0, T < 0\right\} \leq \gamma_n^2/\epsilon^2 + \delta_1(\epsilon).$$

In a similar manner, we can show that

$$\Pr\left\{T_n < 0, T \geq 0\right\} \leq \gamma_n^2/\epsilon^2 + \delta_2(\epsilon),$$

where $\delta_2(\epsilon) = \Pr\left\{\epsilon > T(x_n) \geq 0\right\}$. Then, setting $\delta(\epsilon) = \delta_1(\epsilon) + \delta_2(\epsilon) = \Pr\left\{|T(x)| < \epsilon\right\}$, the difference in risks is dominated by

$$0 \leq R(t_n, \pi) - R(t_B, \pi) \leq a_1\left(\frac{2\gamma_n^2}{\epsilon^2} + \delta(\epsilon)\right). \qquad (1.2.26)$$

To show that $\delta(\epsilon)$ can be made arbitrarily small, it is sufficient to assume that the density functions $f_o(x)$ and $f_1(x)$ are linearly independent, and in addition, they are analytic functions of x.[1]

The linear independence assumption is not unreasonable since, if the density functions were linearly dependent, one could not distinguish between the two hypotheses. The analytic assumption is more then we need, but in the cases we are interested in this assumption will always be met; $f_i(x)$ will be the convolution of a gaussian density (which is analytic) with some distribution function.

These two assumptions imply that the roots of $T(x)=0$ are isolated. For if $T(x)=0$ in some interval then, since $T(x)$ is analytic, $T(x)$ is identically equal to zero. This violates the linear independence assumption. Now, since $T(x)$ is continous, it follows that for any specified δ, we can choose an ϵ such that the probability of the set $A=A\left\{x:|T(x)|<\epsilon\right\}$ satisfies $Pr(A) \leq \delta$. This gives the desired result $Pr\left\{|T(x)|<\epsilon\right\} \leq \delta(\epsilon)$.

We collect our results (and assumptions) in

Theorem 1.2.2: We observe the sequence of stationary dependent random variables X_ℓ, $\ell=1,2,\ldots$, with the marginal density function

$$f(x) = p_o f_o(x) + p_1 f_1(x).$$

Assume that a sequence of test functions, $T_n(x_1,x_2,\ldots x_n;x_n)$ exists which satisfies (22),

---

[1] By linear independence we mean that there does not exist two non-zero constant c_o, c_1, such that $c_o f_o(x) + c_1 f_1(x) = 0$, a.e.x. (Since the f_i are densities, linear dependence is equivalent to equality.)

$$E_n(|T_n(X_1 \ldots X_n; X_n) - T(X_n)|^2) \leq \gamma_n^2,$$

and $\lim_{n\to\infty} \gamma_n^2 = 0$. Define the sequence of decision functions by

$$t_n(x_1, \ldots x_n) = 1 \text{ if } T_n(x_1, x_2, \ldots, x_n; x_n) \geq 0$$

$$= 0 \text{ otherwise.}$$

If $p_i f_i(x) \neq 0$, for a.e.x, i=0,1, the empirical risk converges to the risk incurred using the one-stage procedure. In addition, if the density functions $f_i(x)$, i=0,1, are linearly independent and analytic functions of x, the difference in risks at the n-th decision is bounded by

$$0 \leq R(t_n, \pi) - R(t_B, \pi) \leq a_1\left(2\frac{\gamma_n^2}{\epsilon^2} + \delta(\epsilon)\right).$$

ϵ is an arbitrary positive constant and $\delta(\epsilon)$ can be made arbitrarily small by a suitable choice of ϵ.

We will have the occasion to consider a test function defined as

$$\tilde{T}(x) = s(x)T(x) = s(x)\left\{p_1 a_1 f_1(x) - a_0 p_0 f_0(x)\right\}. \qquad (1.2.27)$$

Since s(x) will be a positive function, the decision function

$$\tilde{t}_B(x) = 1 \text{ if } \tilde{T}(x) \geq 0 \qquad (1.2.28)$$

$$= 0 \text{ otherwise}$$

is identical to $\tilde{t}_B(x)$. Hence, the risk using this equivalent test, $R(\tilde{t}_B, \pi)$, is equal to $R(t_B, \pi)$.

For the empirical procedure, we will then take

$$\tilde{T}_n(x_n) = s(x_n)T_n(x_n) \tag{1.2.29}$$

as the estimate of $\tilde{T}(x)$ and define the decision function by

$$\tilde{t}_n(x_n) = 1 \text{ if } s(x_n)T_n(x_n) \geq 0$$

$$= 0 \text{ otherwise.} \tag{1.2.30}$$

Letting $R(\tilde{t}_n, \pi)$ denote the risk of this procedure, we have

$$R(\tilde{t}_n, \pi) = a_1 p_1 - E_{n\lambda n}(\tilde{t}_n(X_n)b(\lambda_n)). \tag{1.2.21}$$

The difference in risks is

$$0 \leq R(\tilde{t}_n, \pi) - R(\tilde{t}_B, \pi) = E_{n\lambda n}\left\{b(\lambda_n)(\tilde{t}_B(X_n) - \tilde{t}_n(X_n))\right\}$$

$$\leq a_1 E_n\left\{|\tilde{t}_B(X_n) - \tilde{t}_n(X_n)|\right\}. \tag{1.2.32}$$

Then, by a proof which is identical to the previous theorem, we have

Corollary 1.2.2: Assume that the sequence of test functions satisfies

$$E_n\left\{s^2(X_n)(T_n(X_1 \ldots X_n; X_n) - T(X_n))^2\right\} \leq (\gamma_n')^2 \tag{1.2.33}$$

and that $\lim_{n\to\infty} \gamma_n' = 0$. Assume further that the density functions $f_i(x)$, $i=0,1$, are linearly independent and that the functions $s(x)f_i(x)$, $i=0,1$, are analytic functions of x. Then, the difference in risks at the n-th stage is bounded by

$$0 \leq R(\tilde{t}_n, \pi) - R(\tilde{t}_B, \pi) \leq a_1\left(\frac{2\gamma_n'^2}{\epsilon^2} + \delta'(\epsilon)\right) \tag{1.2.34}$$

where $\delta'(\epsilon) = \Pr\{|s(x)T(x)|<\epsilon\}$ can be made arbitrarily small by suitable choice of ϵ.

Similar results can be obtained for any equivalent test procedure. By an equivalent procedure we mean a test function $T_e(x)$ such that for every x, $T_e(x) \geq 0$ when $T(x) \geq 0$, and a decision function $t(x)$ which equals one when $T_e(x) \geq 0$ and zero otherwise. In view of (32), a bound on the difference in risks, analogous to (34), can easily be obtained.

These remarks can be extended to the results of the next two subsections. Since the extension to equivalent tests is straightforward, we will not discuss them further in this chapter.

1.2c Extension Of The Dependent Case To Multiple Hypotheses

Let the parameter λ take on the "values" $\lambda = \{\lambda_0, \lambda_1, \ldots \lambda_k\}$. Again, λ designates which hypothesis is active. We take p_i as the a priori probability of the i-th hypothesis, $\sum p_i = 1$, and $f_i(x)$ as the density function of the observation given that $\lambda = \lambda_i$.

A test procedure is equivalent to specifying (K+1) decision functions $t_i(x)$, $i = 0,1,\ldots K+1$, defined on the space of observations such that if, for a given x, $t_i(x) = 1$ we announce λ_i and if $t_i(x) = 0$, we do not announce λ_i. Clearly, we have $\sum t_i(x) = 1$, for all x.

If we take the loss as 0 for a correct classification and equal to 1 if we are in error then, assuming that $\lambda = \lambda j$, the loss is $(1-(t_j$

$(x))$.[1] The expected loss, given that $\lambda=\lambda_j$, is

$$R(t_j,\lambda) = \int_{-\infty}^{+\infty} (1-t_j(x))f_j(x)dx \qquad (1.2.35)$$

and the overall loss, or expected risk, is

$$R(t,\pi) = \int_{\lambda}\int_{-\infty}^{+\infty} (1-t_j(x))f_j(x)dx\, d\pi(\lambda)$$

$$= \sum_{j=o}^{K} \int_{-\infty}^{+\infty} p_j(1-t_j(x))f_j(x)dx$$

$$= 1 - \sum_{j=o}^{K} p_j \int_{-\infty}^{+\infty} t_j(x)f_j(x)dx. \qquad (1.2.36)$$

We prefer to write this as

$$R(t,\pi) = 1 - p_o - \sum_{j=o}^{K} \int_{-\infty}^{+\infty} t_j(x)T_j(x)dx, \qquad (1.2.37)$$

where the test functions $T_j(x)$ are defined by

$$T_j(x) = p_jf_j(x) - p_of_o(x),\ j=0,1,\ldots K. \qquad (1.2.38)$$

The test procedure given by

$$t_{iB}(x) = 1 \text{ if } T_i(x) \geq T_j(x), \text{ all } j \qquad (1.2.39)$$

$$= 0 \text{ otherwise,}$$

---

[1] We can think of $t_j(x)$ as the probability of announcing $\lambda=\lambda_j$ when we observe x. $(1-t_j(x))$ is then the probability of an error given that $\lambda=\lambda_j$.

minimizes the probability of error. The minimum probability of error is then

$$R(t_B,\pi) = 1 - p_o - \sum_{j=o}^{K} \int_x t_j(x)T_j(x)dx,$$

$$= 1 - p_o - \sum_{j=o}^{K} \int_{A_j} T_j(x)dx,$$

where $A_j = A_j\left\{x: T_j(x) \geq T_i(x),\ i=o,1,\ldots K\right\}$.

When the test functions are unknown, we suppose that we can find a sequence of functions $T_{jn}(x_n)$ which satisfy

$$E_n(|T_{jn}(X_1\ldots X_{n-1},X_n;X_n) - T_j(X_n)|^2) \leq \gamma_{jn}^2, j=o,1,\ldots K \qquad (1.2.40)$$

We then define the sequence of decision functions:

$$t_{jn}(x_1\ldots x_n) = 1 \text{ if } T_{jn}(x_1\ldots x_n;x_n) \geq T_{in}(x_1\ldots x_n;x_n), \text{ all } i,$$

$$= 0 \text{ otherwise.} \qquad (1.2.41)$$

With $f_j(x_1,x_2,\ldots,x_n)$ denoting the joint density function of the n observations given that the n-th λ_n is $\lambda_n=\lambda_{jn}$, the expected risk is

$$R(t_n,\pi) = 1 - \sum_{j=o}^{K} p_j \int\ldots\int t_{jn}(x_1\ldots x_n)f_j(x_1\ldots x_n)dx_1..dx_n.$$

The difference in risks is expressed as

$$0 \leq R(t_n,\pi) - R(t_B,\pi) = \sum_{j=o}^{K} p_j \int\ldots\int (t_{Bj}(x_n) - t_{jn}(x_1\ldots x_n)).$$

$$f_j(x_1\ldots x_n)dx_1\ldots dx_n \qquad (1.2.42)$$

Clearly, $p_j f_j(x_1,\ldots,x_n) \leq \sum_{i=o}^{K} p_i f_i(x_1,\ldots x_n) = f(x_1,\ldots x_n)$, and hence, (42) is dominated by

$$0 \leq R(t_n,\pi) - R(t_B,\pi) \leq \sum_{j=o}^{K} E_n|t_{Bj}(X_n) - t_{nj}(X_1\ldots X_n)| \quad (1.2.43)$$

Let the subscript i in the following expressions read "for all i," and the subscript k mean "for some k." The joint event $((T_{jn}(x_1\ldots x_n; x_n) \geq T_{in}(x_1\ldots x_n;x_n)$ for all i), $T_j(x_n) < T_k(x_n)$ for some k) is written as $(T_{jn} \geq T_{in},\ T_j < T_k)$.

The expectation inside the summation of (43) becomes

$$E_n|t_{Bj}(X_n) - t_{jn}(X_1\ldots X_n)| = \Pr\left\{T_{jn} \geq T_{in}, T_j < T_k\right\} + \Pr\left\{T_{jn} < T_{kn}, T_j \geq T_i\right\}.$$

Consider the first probability expression on the right side. Since the event $(T_{jn} \geq T_{in}, T_j < T_k)$ is included in the event $(T_{jn} \geq T_{kn},\ T_j < T_k)$, the first probability expression is dominated by

$$\Pr\left\{T_{jn} \geq T_{in}, T_j < T_k\right\} \leq \Pr\left\{T_{jn} \geq T_{kn}, T_j < T_k\right\}$$

$$= \Pr\left\{T_{jn} - T_{kn} \geq 0, T_j - T_k \leq -\epsilon_{jk}\right\}$$

$$+ \Pr\left\{T_{jn} - T_{kn} \geq 0,\ -\epsilon_{jk} < T_j - T_k < 0\right\},$$

where ϵ_{jk} is an arbitrary constant. If $(T_{jn} - T_{kn} \geq 0)$ and $((T_j - T_k) \leq -\epsilon_{jk})$ then the expression $(|T_{jn} - T_{kn} - T_j + T_k| \geq \epsilon_{jk})$ holds. Therefore, it follows that

$$\Pr\left\{T_{jn} - T_{kn} \geq 0,\ T_j - T_k \leq -\epsilon_{jk}\right\}$$

$$\leq \Pr\left\{|T_{jn} - T_{kn} - T_j + T_k| \geq \epsilon_{jk}\right\}.$$

Defining $\delta_{1jk} = \Pr\left\{-\epsilon_{jk} < T_j - T_k < 0\right\}$, we have $\Pr\left\{T_{jn} - T_{kn} \geq 0,\ -\epsilon_{jk} < T_j - T_k < 0\right\}$ dominated by δ_{1jk} in analogy to the previous development. We then have the bound

$$\Pr\left\{T_{jn} \geq T_{in}, T_j < T_k\right\}$$

$$\leq \Pr\left\{|T_{jn} - T_{kn} - T_j + T_k| \geq \epsilon_{jk}\right\} + \delta_{1jk}$$

which, in view of (40) and the Minkowski inequality, can be dominated by

$$\Pr\left\{T_{jn} \geq T_{in}, T_j < T_k\right\} \leq \frac{(\gamma_{jn} + \gamma_{kn})^2}{\epsilon_{jk}^2} + \delta_{1jk}. \qquad (1.2.44)$$

Similarly, we can show that

$$\Pr\left\{T_{jn} < T_{kn}, T_j \geq T_i\right\} \leq \frac{(\gamma_{jn} + \gamma_{kn})^2}{\epsilon_{jk}^2} + \delta_{1jk}.$$

Combining these bounds, we obtain

$$E_n|t_{Bj}(X_n) - t_{jn}(X_1, \ldots X_n)| \leq \frac{2(\gamma_{jn} + \gamma_{kn})^2}{\epsilon_{jk}^2} + \delta_{jk}(\epsilon_{jk}),$$

where

$$\delta_{jk} = \Pr\left\{|T_j(x_n) - T_k(x_n)| < \epsilon_{jk}\right\}.$$

In analogy to Theorem 1.2.2, we have

Corollary 1.2.3: We observe the sequence of stationary dependent random variables $X_\ell, \ell = 1,2,\ldots$, with the marginal density function $f(x) = \sum_{j=0}^{k} p_j f_j(x)$. Assume that the sequences of test functions $\{T_{jn}(x_1,x_2,\ldots x_n;x_n)\}$, $j=0,1,\ldots K$, satisfy

$$E_n(T_{jn} - T_j)^2 \leq \gamma_{jn}^2, \quad j=0,1,\ldots K. \tag{1.2.45}$$

and that $\lim_{n\to\infty} \gamma_{jn}=0$, $j=0,1,\ldots K$.

At the n-th decision, define the (K+1) decision functions by

$$t_j(x_1\ldots,x_n) = 1 \text{ if } T_{jn}(x_1\ldots x_n;x_n) \geq T_{in}(x_1\ldots x_n;x_n), \text{ all } i$$

$$= 0 \text{ otherwise, } j=0,1,2,\ldots K.$$

Then, if $p_i f_i(x) \neq$ a.e.x, $i=0,1\ldots K$, the empirical risk converges to the risk of the one-stage procedure. If the density functions $f_j(x)$, $j=0,1,\ldots K$, are linearly independent and analytic functions of x, the difference in risks at the n-th decision is bounded by

$$0 \leq R(t_n,\pi) - R(t_B,\pi) \leq \sum_{j=0}^{K} \left\{ \frac{2(\gamma_{jn} + \gamma_{kn})^2}{\epsilon_{jk}^2} + \delta_{jk}(\epsilon_{jk}) \right\}$$

where again δ_{jk} can be made arbitrarily small.

1.2d Convergence of The Empirical Procedure For Unbounded Loss Functions

The fact that the loss function ((2)) is bounded has been used to considerable advantage in obtaining the above bounds. Situations where $L(t(x),\lambda)$ may not be a bounded function of λ occur when we let λ(the state of nature) take on a continuum of values. We assume that λ is a

random variable drawn from some general papameter space Λ. With λ distributed according to the distribution $\alpha(\lambda)$, the density function of the observation is written

$$f(x) = \int f_\lambda(x)\, d\alpha(\lambda). \tag{1.2.47}$$

The hypothesis test we consider is one in which we infer from the observation X whether $\lambda \epsilon A$ (hypothesis H_o) or $\lambda \epsilon \Lambda - A$ (hypothesis H_1).

To obtain the (one-stage) test procedure, we again let $t(x) = 0$, 1, depending on whether we believe H_o or H_1 is in effect. Defining

$$b(\lambda) = L(0,\lambda) - L(1,\lambda) \tag{1.2.48}$$

and

$$T(x) = \int_\Lambda b(\lambda) f_\lambda(x)\, d\alpha(\lambda), \tag{1.2.49}$$

the risk incurred is a minimum if we choose

$$t_B(x) = 1 \text{ if } T(x) \geq 0$$

$$= 0 \text{ otherwise.}^1$$

The risk is then given by

$$R(t_B,\alpha) = \int_\Lambda L(0,\lambda)\, d\alpha(\lambda) - \int T(x)^+ dx \tag{1.2.50}$$

---

[1] See Robbins [30], section 3, for the details.

which we can also write as

$$R(t_B,\alpha) = \int_{\Lambda} L(0,\lambda)\,d\alpha(\lambda) - E_{x\lambda}\left\{t_B(X)\,b(\lambda)\right\}. \qquad (1.2.51)$$

When the test function $T(x)$ is not known, we define a (two-valued) decision function $t_n(x_1 \ldots x_n)$ as before. The overall (empirical) risk is

$$R(t_n,\alpha) = \int_{\Lambda} L(0,\lambda)\,d\alpha(\lambda) - E_{n\lambda_n}\left\{t_n(X_1 \ldots X_n)\,b(\lambda_n)\right\} \qquad (1.2.52)$$

and the difference in risks can be written as

$$0 \leq R(t_n,\alpha) - R(t_B,\alpha)$$

$$= E_{n\lambda_n}\left\{b(\lambda_n)(t_B(X_n) - t_n(X_n))\right\}. \qquad (1.2.53)$$

If we assume

$$\int b^2(\lambda)\,d\alpha(\lambda) \leq c < \infty \qquad (1.2.54)$$

then by the Schwarz inequality it follows that

$$0 \leq R(t_n,\alpha) - R(t_B,\alpha)$$

$$\leq \left(c\,E_n\left\{|t_B(X_n) - t_n(X_1 \ldots X_n)|^2\right\}\right)^{1/2}. \qquad (1.2.55)$$

The value of the expectation in (55) is identical to the value of expectation appearing in (25). Hence, we obtain

Corollary 1.2.4: Assume there exists a sequence of test functions which satisfies

$$E_n\left\{\left[(T_m(X_1 \ldots X_n; X_n) - T(X_n))^2\right]\right\} \leq \gamma_n^2 .$$

Then, if the decision function $t_n(x_n)$ is defined as

$$t_n(x_n) = 1 \text{ if } T_n(x_n) \geq 0$$

$$= 0 \text{ otherwise,}$$

and if (54) holds, the difference in risks at the n-th stage is dominated by

$$0 \leq R(t_n,\alpha) - R(t_B,\alpha)$$

$$\leq c^{1/2}\left(\frac{2\gamma_n^2}{\epsilon^2} + \delta(\epsilon)\right)^{1/2}. \qquad (1.2.56)$$

Observe that the bound is of order γ_n while the previous bounds on the risk were of order γ_m^2. This is a direct result of the boundedness of $b(\lambda)$ for the minimum probability of error criterion and the fact that the sequence $\{\lambda_\ell\}$ may be dependent.

We have assumed throughout that each decision is based on a single observation. The extension of the above results to more than one sample per decision is straightforward.

1.3 LITERATURE SURVEY AND SCOPE OF THE PRESENT STUDY

We have investigated the convergence of a particular empirical procedure to what we have called the optimum one-stage procedure. By dominating the mean-square error,

$$E_n \quad (\hat{T}_n(X_1 \ldots X_n; X_n) - T(X_n))^2 \quad \leq \gamma_n^2,$$

we are able to bound the rate of convergence of the empirical risk. Hence, the central problem is to find a sequence of estimates $\hat{T}_n(x_1, \ldots x_n; x_n)$ which is consistent, i.e., $\lim_{n\to\infty} \gamma_n = 0$. This is our major concern.

Consider the two hypotheses problem with a minimum probability of error criterion. For this case, $T(x) = p_1 f_1(x) - p_o f_o(x)$. Assume that p_o is known and that the densities $f_o(x)$ and $f_1(x)$ are unknown. To estimate $T(x)$, a natural procedure would be to first estimate the densities and then take

$$\hat{T}_n(x) = p_1 \hat{f}_1(x) - p_o \hat{f}_{on}(x) \qquad (1.3.1)$$

as the estimate of the test function for the n-th decision. If $\hat{f}_{1n}$ and $\hat{f}_{on}$ are consistent estimates then the sequence $\hat{T}_n(x)$ will also be consistent. The manner in which the estimates are obtained depends on whether "learning" samples are available.

If one can classify an observation with probability one, it is called a learning sample. Then, if the observation is known to come from, say, hypothesis H_o, we would use it to update our estimate of $f_o(x)$. This type of operation has sometimes been called supervised learning or learning with a teacher.[1]

---

[1] In the context of communication problems, learning samples are provided by periodically injecting a known fixed sequence into the sequence of information bearing signals, i.e., channel sounding signals. See [35,36]. Learning samples of a different nature occur in problems such as statistical weather prediction. Based on some observational data, an inference is made about the future weather. At some later time we find out if the inference was correct. This knowledge would then be used to form better inference procedures.

When learning samples are not available, the problem is more difficult. Since we never know from which population the observation is drawn, we can not directly estimate the desired quantities. One possible procedure is to estimate the overall density function, $f(x) = p_0 f_0(x) + p_1 f_1(x)$, and then attempt to extract from this estimate the parts that are unknown and that are needed to form the estimate of the test function. This mode of operation has been called nonsupervised learning or learning without a teacher. We remark that learning in the nonsupervised mode is not always possible.

When the sequence of observations is independent and if, with either of the above procedures, we obtain consistent estimates of the test function, then, these procedures are asymptotically optimal.[2] This is not to say that the probability of error is minimized at each stage. This, of course, depends on what part of $f_i(x)$ is unknown, how it is estimated and subsequently used to form the estimate of the test function.

1.3a Literature Survey

The learning procedures most frequently investigated are those in which a set of parameter vectors, $\underline{\theta}_i$, $i=1,\ldots k$, is to be estimated

[2] When the observations are dependent, the procedures are in no way optimum. Presumably, they are reasonable procedures to follow, especially when the exact nature of the dependency on the observations in not specified.

from the statistically related observation vectors $\underline{X}_\ell, \ell=1,\ldots n.$[1] Each parameter (or pattern class) $\underline{\Theta}_i$ is associated with a particular hypothesis H_i and could represent samples of a signal which is buried in noise. The density function of the observation given that H_i is active is written as $f_{\Theta_i}(\underline{x}_i)$, and the overall density function of the observation becomes

$$f(\underline{x}_i) = \sum_{i=1}^{K} p_{\underline{\Theta}_i} f_{\underline{\Theta}_i}(\underline{x}). \qquad (1.3.2)$$

Furthermore, it is assumed that the set of patterns is initially chosen from a known a prior distribution, $p_{\underline{\Theta}_i}(\ )$, and then held fixed for the experiment. The statistical inference problem is to decide which hypothesis (pattern class) is in effect for a particular observation $\underline{X}$. The criterion used is the minimization of the total probability of error.

Within this framework, a number of authors (e.g., [1,5,20,21]) have investigated optimum test procedures when learning samples are available. Let χ_k represent the sequence of learning samples. Then, given the observation $\underline{X}$, the optimum decision rule is to compute the a posteriori conditional densities

---

[1] Vectors are denoted by the _ notation.

$$P\left\{\underline{\theta}_j \mid \underline{X}, \chi_k\right\} \qquad ,j=1,\ldots K \qquad (1.3.3)$$

and announce the $\underline{\theta}_j$ for which (3) is a maximum.

Braverman[5], assumes that the sequence of learning samples $\chi_k = \underline{X}_{k_1}, \underline{X}_{k_2}, \ldots$ is independent and that the learning samples of one class impart no information concerning the patterns of another class. Letting χ_{k_j} denote the set of learning samples of the j-th class, (3) becomes

$$P\left\{\underline{\theta}_j \mid \underline{X}, \chi_{k_j}\right\} \qquad ,j=1,\ldots K.$$

He takes the density function $f_{\underline{\theta}_j}(\underline{x})$ as gaussian ($f_{\underline{\theta}_j}(\underline{x}) = g(\underline{x}-\underline{\theta}_j)$) and the a priori densities $p_{\underline{\theta}_j}(\ )$, $j=1,\ldots K$. also as gaussian with unknown means and known covariance. The optimum procedure is then to use the learning samples to estimate the means of each class and use these estimates in the computation of the a posteriori probabilities. For the case of two hypotheses, he shows that the difference between the error probability of the above procedure and the error probability in the case the patterns (mean values) are known is approximately inversely proportional to the number of learning samples.

Keehn[21], extends the work of Braverman by taking both the mean vector and covariance matrix of $P_{\underline{\theta}_j}(\ )$ as unknown.

Scudder[39,40], also takes the noise and a priori distributions as independent gaussian and investigates the problems encountered when learning samples are not available. The optimum test procedure now

requires an exponentially growing memory. He then looks at a fixed memory technique similar to the procedure used when learning samples are available, but now, learning takes place on the basis of previous decisions which are never known with certainty to be correct.

The problem of when the optimum test procedure, with or without learning samples, requires a growing memory is discussed in a paper by Spragins[42]. The optimum test procedure (an application of Bayes' rule conditioned on an increasing number of observations) will be of fixed memory if and only if the sequence of (independent) observations admits a sufficient statistic of fixed dimension. The existence of the sufficient statistic is seen to imply the existence of an a priori distribution $P_{\underline{\theta}_j}(\ )$ which has a "reproducing" property. Thus, by choosing an a priori distribution which has the reproducing property, a number of authors (e.g. [5,21]) are able to obtain optimum fixed memory procedures.

Hancock and Patrick[17] provide for a general formulation of the learning problem by focusing attention on the overall distribution as given by (2). An important contribution of this study is the determination of when sufficient amounts of a priori information exists for a learning procedure to converge. When little a priori information is known, they apply histogram techniques to a class of nonsupervisory problems. When the functional form of the overall density is known, they investigate estimates of the parameters $\underline{\theta}_j$ which characterize the

overall distribution or, as they call it, the mixture. The estimates are shown to be consistent thus leading to an asymptotically optimal test procedure.

Somewhat related, but less general in formulation, is the work of Cooper and Cooper[9]. They consider the two-category problem with particular emphasis on the case where the overall density is the sum of two gaussian densities. Taking each hypothesis equiprobable, they discuss different estimates of the unknown means which are then used to form an estimate of the test function. They extend the (nonsupervisory) results to multivariate gaussian densities by estimating the parameters which characterize the optimum partition (i.e., a hyperplane) of the sample space. Also discussed is the case where the arbitrary densities of the two equiprobable hypotheses differ only in a location parameter.

A departure in the above formulation is made by Robbins[29-31] and his associates [19,37], They consider only one a priori distribution, $p_\theta(\ )=p(\theta)$, and take the distribution as unknown. Here, the inference problem is to decide whether θ is contained is some set A or its complement. Since the density function of an observation under either hypothesis is the same, $f(x)=\int f_\theta(x)\,dp(\theta)$, every observation can be considered a learning sample even though these observations are never classified correctly with probability one. Their main effort is directed toward showing that the empirical procedures are asymptotically optimal

for a variety of hypotheis testing (and estimation) problems.

All of the above authors take sequences of independent observations. Tainiter[44] extends one aspect of the work of Robbins to M-dependent observations and Raviv[28] takes the "patterns" to be a Markov sequence with the transition probability matrix initially unknown.

Special formulations and learning procedures appropriate to certain communication problems are given by Glaser[15], Price and Green[27] and Sebestyen[41]. A bibliography emphasising the supervised mode of learning is given in [2]. A discussion of most of the approaches to non-supervised learning is given in the recent paper by Spragins[43].

1.3b Scope of the Present Study

The present study is closest, in spirit, to the work of Robbins. The problems we will consider are those in which the "patterns" are random variables. Thus, if the same pattern class or hypothesis is active in succeeding intervals, this only means that the distributions from which they are drawn are the same. It is these a priori distributions which we will take as unknown.

In particular, we shall consider a class of problems where the marginal density function of a single observation can be written as

$$f(x) = \int g(x-z;\sigma)\, d\alpha(z), \qquad (1.3.4)$$

with a corresponding vector equation for multidimensional observations. $g(x;\sigma)$ denotes the gaussian density function with standard deviation σ.

By suitably interpreting $\alpha(z)$, we can include all the problems we are interested in. We give the following as examples.

Let $u(t)$ be the unit step function and define

$$\alpha(z) = \sum_{i=1}^{K} p_i u(z-y_i). \tag{1.3.5}$$

Then, $f(x)$ becomes

$$f(x) = \sum_{i=1}^{K} p_i g(x-y_i;\sigma). \tag{1.3.6}$$

This represents the density function of the observation where one of K signals is transmitted with gaussian noise added to the message. The signals represent the values the random variable can assume.

A generalization of (5) is to take

$$\alpha(z) = \sum_{i=1}^{K} p_i \int u(z-y)\, d\beta_i(y) \tag{1.3.7}$$

where $\beta_i(y)$ represents one of K different distributions. $f(x)$ is then given by

$$f(x) = \sum_{i=1}^{K} p_i \int g(x-y;\sigma)\, d\beta_i(y). \tag{1.3.8}$$

Here the problem would be one of testing between K composite hypotheses with noise-like signals.

Letting $u(z-y) = u(z-s_i(t,\underline{y}))$ in (7) gives

$$f(x) = \sum_{i=1}^{K} p_i \int g(x-s_i(t,\underline{y});\sigma)\, d\beta_i(y) \tag{1.3.9}$$

This has the interpretation as the overall density function of K composite hypotheses with the i-th hypothesis representing the s_i signal being transmitted. The a priori probability of this transmission is p_i. The notation $s_i(t,\underline{y})$ is taken to mean that the signal s_i which, for example, is time sampled at t, is distorted by the random vector $\underline{y}$.

The difficulties we shall encounter are not in attempting to process the observations is some optimal fashion. We have already agreed to consider a learning procedure which, at best, converges to the optimum one-stage procedure; this empirical procedure being asymptotically optimum if the observations are independent. Our difficulties will stem from the fact that the a priori distribution $\beta(y)$ is taken as completely unknown as opposed to assuming some known functional form with a finite set of unknown parameters.[1]

The empirical procedure we have outlined is one of estimating the densities $f_i(x)$ when learning samples are available, and, initially, the overall density $f(x)=\sum_i p_i f_i(x)$ when operating in the nonsupervisory mode. Much of this study deals with estimating f(x) as given by (4), and establishing bounds on the mean-square error in the estimate.

---

[1] There is one exception. We also consider (6) with the a priori probabilities unknown.

In Chapter 2 we consider different methods of estimating $f(x)$. Of particular interest is an eigenfunction representation (section 2.5) for $f(x)$ which we obtain by solving an eigenfunction problem associated with equation (4). Chapter 3 extends the results to estimating the k-variate density function $f(x_1 \ldots x_k)$.

In Chapter 4, we apply our results to some problems in communication theory. Section 4.1 considers transmission through a general, stationary, random channel when learning samples are available. This problem serves to relate the results of section 1.2 on the convergence of the empirical procedure with our results on density estimation. It also illustrates when we can expect to obtain solutions to the nonsupervisory problem. The remaining applications emphasize learning in the nonsupervisory mode. In section 4.2 we consider the problem of transmission of known signals with unknown a priori probabilities and in section 4.3 we discuss the problem of transmission through a random multiplicative channel. In section 4.4 we consider a problem with an unbounded loss function.

A summary of this study is given in Chapter 5.

CHAPTER 2

ESTIMATING THE DENSITY FUNCTION OF OBSERVATIONS—UNIVARIATE CASE

2.1 INTRODUCTION

As discussed in the previous chapter, one approach to finding a convergent sequence of test functions is to first obtain a convergent sequence of estimates for the unknown density functions. These estimates are then used to form a test function, the structure of which is identical to the test function one would use if all distributions were known. Our main concern in this chapter is obtaining consistent estimates of the univariate density function of the observations. By consistent estimates we will mean estimates which converge in mean-square in the sense of

$$\lim_{n\to\infty} E\{(f(x) - \hat{f}_n(X_1,X_2,\ldots X_n;x))^2\} = 0 \quad \text{for every } x \qquad (2.1.1)$$

or

$$\lim_{n\to\infty} E\{(f(X_n) - \hat{f}_n(X_1,X_2,\ldots X_n;X_n))^2\} = 0. \qquad (2.1.2)$$

Equation (1), obviously, is concerned with convergence to the constant $f(x)$, while in (2) we have convergence to a random variable. It is (2) which we need to demonstrate convergence of the empirical procedure for the case of dependent samples.[1] Since, for two of the methods which we

[1] The convergence in (1) is essentially that required for the case of independent samples. See (1.2.16).

use to estimate $f(x)$, there is little difference between consistency in the sense of (1) and (2), we will evaluate bounds for both types of convergence.

To compare our results with previous work in the area of density estimation, we will also consider a global measure of the error, the mean integrated square error,

$$E\int_{-\infty}^{+\infty}(f(x) - \hat{f}_n(X_1,X_2,\ldots X_n;x))^2\,dx \qquad (2.1.3)$$

The basic problem which we want to discuss is as follows. We are given the stationary sequence of identically distributed random variables (observations), $X_i = N_i + Z_i$, $i=1,2,\ldots$, where N_i is a sample from a stationary gaussian process and Z_i is a sample from an unknown random process. The samples may be time samples or any other linear processing of the received waveform which preserves the gaussian nature of the noise. With N_i independent of Z_i, the univariate density function of the observation X_i is

$$f(x) = \int_{-\infty}^{+\infty} g(x-z;\,\sigma)\,d\alpha(z)\ , \qquad (2.1.4)$$

where by $g(x-y;\sigma)$ we mean the gaussian density function with mean value y and standard deviation σ. We want to take the gaussian noise samples as correlated and also consider a dependency on the Z_i sequence which will be specified later.

In the next section we consider the empirical distribution function as an estimate of the cumulative distribution of the observations. We

investigate the mean-square error and obtain a bound on the rate of convergence. The results of this section are then applied to the problem of estimating the density function, for which we give three techniques.

The first method of estimating $f(x)$, section 2.3, is analogous to the technique used in estimating a spectral density. For this method, we restrict our study of convergence to those as specified by equations (1) and (3). This method of estimation requires a minimum of assumptions to guarantee convergence.

In section 2.4 we consider an orthogonal representation for $f(x)$ and investigate all three of the above modes of convergence.

The method in section 2.5 is analogous to the technique generally used to solve a deterministic integral equation. To the best of our knowledge, this approach has not appeared in the literature.

The results which we will need for the applications of the empirical Bayes procedure are contained in Corollaries 2.4.1, 2.5.1, and section 2.6. Section 2.6 considers a special form of $\alpha(z)$; the case where $\alpha(z)$ contains a finite set of unknown parameters which enter linearly into $f(x)$.

A summary of the chapter and generalizations are given in section 2.7.

2.2 THE EMPIRICAL DISTRIBUTION FUNCTION

We want to consider the empirical distribution function as an estimate of the true distribution. For the case of independent observations, it

is easy to see that this estimate is consistent with the mean-square error going to zero at a rate $1/n$. For the case of dependent samples, our main interest will be a characterization of the nature of the dependency on the samples, or the underlying random process, for which we can still guarantee consistency with a specified rate of convergence.

The sequence of observations, $X_1, X_2, \ldots X_n$ are identically distributed (not necessarily independent) random variables. X_i is a sample from a stationary process which is composed of the sum of a gaussian process with an autocorrelation function $R(t)$, and another stationary processes $Z(t)$; $X_i = N_i + Z_i$. With N_i and Z_i independent, the density function of the observation is given by

$$f(x) = \int_{-\infty}^{+\infty} g(x-z;\sigma)d\alpha(z) , \qquad (2.2.1)$$

with the corresponding distribution function

$$F(x) = \int_{-\infty}^{x} f(y)dy . \qquad (2.2.2)$$

The empirical distribution function of the observations is defined as

$$F_n(x) = \frac{1}{n} (\text{number of } X_i \leq x, \ i=1,2,\ldots n).$$

Let $U_\ell(X_\ell) = 1$ if $X_\ell \leq x$, and equal zero otherwise. $F_n(x)$ is then written as

$$F_n(x) = \frac{1}{n} \sum_{\ell=1}^{n} U_\ell(X_\ell) . \qquad (2.2.3)$$

With E denoting the mathematical expectation, we have

$$E(F_n(x)) = F(x) . \tag{2.2.4}$$

$F_n(x)$ is an unbiased estimate of the distribution whether or not the observations are independent.

The mean-square error can be written in terms of a bias and variance contribution:

$$E(F(x) - F_n(x))^2 = (E(F(x) - F_n(x)))^2 + E(F_n(x) - E(F_n(x)))^2 . \tag{2.2.5}$$

Since the first term is zero to investigate consistency we need only consider the variance of the estimate.

The second moment is given by

$$\begin{aligned} E\{F_n^2(x)\} &= \frac{1}{n^2} E\left\{\sum_{\ell=1}^{n} U_\ell(X_\ell)\right\}^2 \\ &= \frac{1}{n^2} E\left\{\sum_{\ell=1}^{n} U_\ell{}^2(X) + 2\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n} U_\ell(X_\ell)U_m(X_m)\right\} \\ &= \frac{F(x)}{n} + \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n} F_{\ell m}(x,x) . \end{aligned} \tag{2.2.6}$$

We have defined $F_{\ell m}(x,x)$ as the joint probability that the samples from the ℓ and m intervals are less than or equal to x,[1]

[1] The subscripts ℓ and m will always mean the ℓ and m observations (decisions, intervals, etc...) when used in a double sum.

$$F_{\ell m}(x,x) = \Pr\{X_\ell \le x,\ X_m \le x\} = F_{m-\ell}(x,x)\ . \qquad (2.2.7)$$

We want to display the effect of the dependency of the observations on the variance of $F_n(x)$. Add $(1-1/n)F^2(x)$ to (6) and subtract its equivalent

$$\frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n} F^2(x)\ .$$

The variance is

$$V(F_n(x)) = \frac{F(x)}{n}(1 - F(x)) + \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}(F_{m-\ell}(x,x) - F^2(x))\ . \qquad (2.2.8)$$

With independent observations the second expression on the right side of (8) is zero—the variance reduces to the standard result.

Assume, for the moment, that the sequence of random variables Z_i are independent. Then, the second-order distribution is

$$F_{m-\ell}(x,x) = \int_{-\infty}^{x}\int_{-\infty}^{x}\int_{z_1}\int_{z_2} g_2(y_1-z_1, y_2-z_2; \sigma, \rho_{m-\ell})\,d\alpha(z_1)\,d\alpha(z_2)\,dy_1, dy_2, \qquad (2.2.9)$$

where $g_2(n_1,n_2;\sigma,\rho_{m-\ell})$ is the bivariate gaussian density function with the random variables N_ℓ and N_m having the same standard deviation σ and a correlation coefficient $\rho_{m-\ell} = R(m-\ell)/R(0)$.[1] The univariate distribution

[1] If we had time samples, $\rho_{m-\ell}=R((m-\ell)T)/R(0)$, where T is the time between succeeding samples. We shall take T = 1. The gaussian random variables have the same standard deviation since the waveform in each interval is identically processed.

$\alpha(z_i)$ is independent of the subscripts ℓ or m because of the assumed stationarity.

It will be convenient to denote the expression in the double summation of (8) by $\tilde{D}_{m-\ell}$,

$$\tilde{D}_{m-\ell} = F_{m-\ell}(x,x) - F^2(x)$$

$$= \int_{-\infty}^{x}\int_{-\infty}^{x}\int_{z_1}\int_{z_2} [g_2(y_1-z_1,y_2-z_2;\sigma,\rho_{m-\ell})-g(y_1-z_1;\sigma)g(y_2-z_2;\sigma)]d\alpha(z_1)$$

$$d\alpha(z_2)dy_1\, dy_2 \qquad (2.2.10)$$

We interchange the y and z integrations((10) is absolutely integrable) and consider the resulting inner double integrals

$$\int_{-\infty}^{x}\int_{-\infty}^{x} [g_2(y_1-z_1,y_2-z_2;\sigma,\rho_{m-\ell})-g(y_1-z_1;\sigma)g(y_2-z_2;\sigma)]dy_1\, dy_2 \ . \qquad (2.2.11)$$

The bivariate gaussian density is expressed in terms of Mehler's formula. From Appendix A, (A.20), with the requirement that $|\rho_{m-\ell}| < 1$, $\ell \neq m$, we have:

$$g_2(y_1-z_1,y_2-z_2;\sigma,\rho_{m-\ell}) = g(\tfrac{y_1-z_1}{\sigma})g(\tfrac{y_2-z_2}{\sigma}) \sum_{j=0}^{\infty} \frac{\rho_{m-\ell}^j}{j!} He_j(\tfrac{y_1-z_1}{\sigma})\ He_j(\tfrac{y_2-z_2}{\sigma}) \ . \qquad (2.2.12)$$

The $He_j(y/\sigma)$ are the Hermite polynomials orthogonal with respect to the gaussian weight $g(y/\sigma)$.[1] Observe that this is not an orthogonal expansion

[1] We use the notation $g(y/\sigma)$ for the gaussian density (with standard deviation σ) when dealing with the corresponding polynomials. $g(y/\sigma)$ is identical to $g(y;\sigma)$ which is the notation we generally use for the gaussian density.

in the usual sense. The polynomials are defined in such a way that the orthogonal functions are given by $\sqrt{g(y)}\,He_j(y)$.[1]

Substitute (12) for the bivariate density in (11). The first term of the series cancels leaving

$$\sum_{j=1}^{\infty} \frac{\rho_{m-\ell}^{j}}{j!} \int_{-\infty}^{x}\int_{-\infty}^{x} He_j(\tfrac{y_1-z_1}{\sigma})g(\tfrac{y_1-z_1}{\sigma})\; He_j(\tfrac{y_2-z_2}{\sigma})g(\tfrac{y_2-z_2}{\sigma})dy_1\, dy_2 \ . \quad (2.2.13)$$

With $|\rho_{m-\ell}| < 1$, it is easy to justify the above inversion of summation and integrations.[2] The integrals are then dominated using Schwarz's inequality and the orthogonality relation for the Hermite polynomials (A.10):

$$\int_{-\infty}^{x} g(\tfrac{y-z}{\sigma})\; He_j(\tfrac{y-z}{\sigma})dy$$

$$\leq \left[\int_{-\infty}^{+\infty} g(\tfrac{y-z}{\sigma})dy \int_{-\infty}^{+\infty} g(\tfrac{y-z}{\sigma})\; He_j^2(\tfrac{y-z}{\sigma})dy\right]^{1/2}$$

$$\leq \sqrt{j!} \ . \quad (2.2.14)$$

Hence, (13) is bounded by

$$\sum_{j=1}^{\infty} |\rho_{m-\ell}|^{j} = \frac{|\rho_{m-\ell}|}{1 - |\rho_{m-\ell}|} \ .$$

$\tilde{D}_{m-\ell}$ is also bounded by the same quantity,

$$\tilde{D}_{m-\ell} \leq \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{|\rho_{m-\ell}|}{1 - |\rho_{m-\ell}|}\, d\alpha(z_1)d\alpha(z_2) = \frac{|\rho_{m-\ell}|}{1 - |\rho_{m-\ell}|} \ . \quad (2.2.15)$$

---

[1] See Appendix A, section A.1.

[2] See Appendix A, Lemma A.1.

Notice that (15) would still be a valid bound on $\tilde{D}_{m-\ell}$ if the integrand in (10) were replaced by its absolute value. We shall use this later.

Combining (8) and (15), the variance of the empirical distribution function is dominated by

$$V(F_n(x)) = \frac{F(x)}{n}(1-F(x)) + \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}\tilde{D}_{m-\ell}$$

$$\leq \frac{F(x)}{n}(1-F(x)) + \frac{2}{n^2}\sum_{\tau=1}^{n}(n-\tau)\frac{|\rho_\tau|}{1-|\rho_\tau|}, \qquad (2.2.16)$$

with the second expression following from the stationarity and with $\tau=m-\ell$.

Convergence in quadratic mean requires that $V(F_n(x)) \to 0$ as $n \to \infty$. For this, it is sufficient to assume that the autocorrelation function of the gaussian noise process satisfies

$$R(\tau) \to 0 \quad \text{as} \quad \tau \to \infty . \qquad (2.2.17)$$

Condition (17) excludes the possibility of jumps in the spectrum of the noise process. It then follows (Loève [24], p. 202) that

$$\rho_* = \max_{\tau \geq 1} |\rho_\tau| < 1 . \qquad (2.2.18)$$

Using this fact, the second part of (16) is majorized by

$$\frac{2}{n^2}\sum_{\tau=1}^{n}(n-\tau)\frac{|\rho_\tau|}{1-|\rho_\tau|} < \frac{2}{1-\rho_*}\frac{1}{n}\sum_{\tau=1}^{n}|\rho_\tau| . \qquad (2.2.19)$$

Since $\rho_\tau \to 0$, it follows that the sequence of arithmetic means, $\frac{1}{n}\sum_{\tau=1}^{n}|\rho_\tau|$

tends to zero as $n \to \infty$ (Hobson [18], p. 7).

Notice also that (18) is sufficient for the validity of the Mehler formula.

With the first part of (16) dominated by $1/n$, we have

Theorem 2.2.1: Given the sequence of identically distributed random variables of the form $X_i = N_i + Z_i$, with the univariate density function given by

$$f(x) = \int_{-\infty}^{+\infty} g(x-z;\sigma)\, d\alpha(z) .$$

Assume that

i) the sequence of random variables $\{Z_i\}$ is independent

ii) the autocorrelation function of the gaussian noise satisfies

$R(\tau) \to 0$ as $\tau \to \infty$

Then, the empirical distribution function is a consistent estimate of $F(x)$. Upon applying the Chebyshev inequality, and since $V(F_n(x)) \to 0$ uniformly in x, we also have uniform convergence in probability: as $n \to \infty$ and for arbitrary ϵ, $\Pr\{|F_n(x)-F(x)| > \epsilon\} \to 0$, uniformly in x.[1]

In order tc obtain a bound for the rate of convergence we need to specify the manner in which $R(\tau) \to 0$.[2] For example, assume that $R(\tau)$ is bounded by

---

[1] We note that the hypothesis of the theorem is sufficient to ensure convergence with probability one. This will be discussed at the end of the section.

[2] The bounds given below are for time samples, with the time between succeeding samples taken as $T = 1$.

$$|R(\tau)| \leq \sigma^2/\tau^{\delta} \tag{2.2.20}$$

for $|\tau| \geq B_1$ where $0 < \delta < 1$. Then, it is not difficult to obtain an integral upper bound for the arithmetic mean:

$$\frac{1}{n}\sum_{\tau=1}^{n} |\rho_\tau| = \frac{1}{n}\sum_{\tau=1}^{n} \frac{|R(\tau)|}{\sigma^2} < \frac{B_1}{n} + \frac{1}{(1-\delta)n^{\delta}} \quad . \tag{2.2.21}$$

In the sequel, we will designate (20) as condition A.

Alternatively, we could make the assumption that

$$\int_0^{\infty} |R(t)|dt < \infty \ . \tag{2.2.22}$$

This implies that the spectrum is absolutely continuous. Then, by the Riemann-Lebesgue lemma we have $R(t) \to 0$ and $t \to \infty$. Assuming further that $R(t)$ is monotonically decreasing for $|t| > B_1$, an integral upper bound is given by

$$\frac{1}{n}\sum_{\tau=1}^{n} |\rho_\tau| < \frac{B_2}{n} \quad , \tag{2.2.23}$$

where we have set

$$B_2 = B_1 + \frac{1}{\sigma^2}\int_{B_1}^{\infty} |R(t)|dt \ . \tag{2.2.24}$$

The assumption of monotonicity can be dropped if the autocorrelation possesses a derivative which is integrable. Then, we replace B_2 in (23) by

$$B_2 = \frac{1}{\sigma^2}\int_0^{\infty} (|R(t)| + |R'(t)|)dt \ . \tag{2.2.25}$$

This follows from the Euler-Maclaurin summation formula.

In the sequel, equation (23) will be designated as condition B, with the constant B_2 given either by (24) or (25).

Corollary 2.2.1: Under the hypotheses of Theorem 2.2.1 and with $R(t)$ satisfying condition A, from (16) and (19), the variance of the empirical distribution function is dominated by

$$V(F_n(x)) \leq \frac{1}{n} + \frac{2}{1-\rho_*}\left(\frac{B_1}{n} + \frac{1}{(1-\delta)n^{\delta}}\right). \qquad (2.2.26)$$

Alternatively, if $R(\tau)$ satisfies condition B, the variance is dominated by

$$V(F_n(x)) \leq \frac{1}{n}\left(1 + \frac{2B_2}{1-\rho_*}\right). \qquad (2.2.27)$$

Note that the bound in (27) gives the same rate of convergence as in the case of independent samples.

As easy extension of Theorem 2.2.1, and one of practical importance, can be obtained by replacing the independence assumption on the Z random variables by one of M-dependence.

Definition: The random variables Z_ℓ and Z_m are said to be M-dependent if the variables Z_ℓ and Z_m are independent for $|m-\ell| \geq M$. In terms of the distributions, we have

$$\alpha_{m-\ell}(z_1, z_2) = \alpha(z_1)\,\alpha(z_2) \quad \text{for } |m-\ell| > M,$$

where

$$\alpha_{m-\ell}(z_1,z_2) = \Pr\{Z_\ell \le z_1 \ , \ Z_m \le z_2\} \ .$$

The extension is carried out by noting that the independence of Z_ℓ and Z_m was first used in (9). In general, this equation now becomes

$$F_{m-\ell}(x,x) = \int_{-\infty}^{x}\int_{-\infty}^{x}\int_{z_1}\int_{z_2} g_2(y_1-z_1,y_2-z_2;\sigma,\rho_{m-\ell})\,d\alpha_{m-\ell}(z_1,z_2)\,dy_1\,dy_2 \ . \tag{2.2.28}$$

We use (28) in the expression $F_{m-\ell}(x,x)-F^2(x)$, add and subtract (9), and group the terms so as to display the Z dependence. Designating the resulting expression by $D_{m-\ell}$, we obtain

$$\begin{aligned} D_{m-\ell} = {} & \int_{-\infty}^{x}\int_{-\infty}^{x}\int_{z_1}\int_{z_2} \{g_2(y_1-z_1,y_2-z_2;\sigma,\rho_{m-\ell})-g(y_1-z_1;\sigma)g(y_2-z_2;\sigma)\} \\ & \qquad d\alpha(z_1)\,d\alpha(z_2)\,dy_1\,dy_2 \\ & + \int_{-\infty}^{x}\int_{-\infty}^{x}\int_{z_1}\int_{z_2} g_2(y_1-z_1,y_2-z_2;\sigma,\rho_{m-\ell})\{d\alpha_{m-\ell}(z_1,z_2)-d\alpha(z_1)\,d\alpha(z_2)\} \\ & \qquad dy_1\,dy_2 \end{aligned} \tag{2.2.29}$$

The first term on the right is the same as before and is bounded by (15). The second expression is easily dominated by:

$$D_{m-\ell} \le 2 \qquad \text{for } |\ell-m| < M$$

and

$$D_{m-\ell} = 0 \qquad \text{for } |\ell-m| \ge M$$

Using these bounds and the previous results we have

Corollary 2.2.2: Given the hypotheses of Theorem 2.2.1 but with condition i) replaced by one of M-dependence. Then, the variance satisfies

$$\lim_{n\to\infty} E(F_n(x) - F(x))^2 = \lim_{n\to\infty} V(F_n(x)) = 0,$$

uniformly in x. In addition, if R(t) satisfies condition A, we have the bound

$$V(F_n(x)) \leq \frac{1+4(M-1)}{n} + \frac{2}{1-\rho_*}\left(\frac{B_1}{n} + \frac{1}{(1-\delta)n^{\delta}}\right) . \qquad (2.2.30)$$

With R(t) satisfying condition B, the variance is dominated by

$$V(F_n(x)) \leq \frac{1}{n}\left(1 + 4(M-1) + \frac{2\,B_2}{1-\rho_*}\right) . \qquad (2.2.31)$$

We now replace the M-dependence assumption with an ergodic requirement.

Suppose that the stationary sequence $\{Z_\ell\}$ is ergodic. Now, the weakest condition we have imposed on the correlation function of the gaussian process ($R(t) \to 0$ as $t \to \infty$) implies that the spectrum of the process is continuous. This, in turn, is a known necessary and sufficient condition for the gaussian process to be ergodic.[1] Since N_i and Z_i are independent it follows that $X_i = N_i + Z_i$ is an ergodic sequence.

We have previously defined the random variable U_ℓ as: $U_\ell(X_\ell) = 1$ if $X_\ell \leq x$, and = 0 otherwise. Since $F_n(x) = \frac{1}{n}\sum_{\ell=1}^{n} U_\ell(X_\ell)$ and $E(U_\ell(X_\ell))^2 = E(U_\ell(X_\ell)) = F(x) \leq 1$, we can use the Mean Ergodic Theorem([16], p. 16) to get:

$$\lim_{n\to\infty} E(|F_n(X_1,X_2,\ldots X_n;x) - \tilde{F}(x)|^2) = 0 \qquad (2.2.32)$$

for every x. Hence, we always have mean-square convergence to some $\tilde{F}(x)$. In addition, with the $\{X_\ell\}$ sequence ergodic, we have from Birkhoff's ergodic

[1] Grenander, U., "Stochastic Processes and Statistical Inference," Arkiv fur Matematik, vol. 17, 1950, pp. 195-277.

theorem ([16], p. 18):

$$\lim_{n\to\infty} F_n(X_1,X_2,\ldots,X_n;x) = E(F_n(X_1\ldots X_n;x) = F(x) \qquad (2.2.33)$$

with probability one. Since $F_n(x)$ converges with probability one to $F(x)$ and in mean-square to $\tilde{F}(x)$, it follows that $\tilde{F}(x) = F(x)$ with probability one for every x. Thus, convergence of $F_n(x)$ to the true distribution function is ensured under an ergodic condition on the $\{Z_\ell\}$ sequence and the above condition on the autocorrelation function of the gaussian process. What we do not have is a measure of how fast the convergence takes place. We now want to find what conditions are required to characterize a rate of convergence. In doing this, we will also directly verify that $\tilde{F}(x) = F(x)$ when the Z process is ergodic.

Consider the expression for $V(F_n(x))$ in the case the Z_i are dependent variables:

$$V(F_n(x)) = \frac{F(x)}{n}(1 - F(x)) + \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n} D_{m-\ell}$$

$$= \frac{F(x)}{n}(1 - F(x)) + \frac{2}{n^2}\sum_{\tau=1}^{n}(n-\tau)\,D_\tau \qquad (2.2.16)$$

From equation (29), D_τ is the sum of two terms. The first part of D_τ has already been bounded by (15):

$$V(F_n(x)) \le \frac{F(x)}{n}(1-F(x)) + \frac{2}{n^2}\sum_{\tau=1}^{n}(n-\tau)\,\frac{|\rho_\tau|}{1 - |\rho_\tau|}$$

$$+\frac{2}{n^2}\sum_{\tau=1}^{n}(n-\tau)\,D_{\tau,2}\,. \tag{2.2.34}$$

$D_{\tau,2}$ designates the second term of D_τ,

$$D_{\tau,2}=\int_{-\infty}^{x}\int_{-\infty}^{x}\int_{z_1}\int_{z_2} g_2(y_1-z_1,y_2-z_2;\sigma,\rho_\tau)[d\alpha_\tau(z_1,z_2)-d\alpha(z_1)d\alpha(z_2)]dy_1dy_2\,. \tag{2.2.35}$$

Perform the y integrations first and let G_2 denote the second-order gaussian distribution function.

$$D_{\tau,2}=\int_{z_1}\int_{z_2}[G_2(x-z_1,x-z_2;\sigma,\rho_\tau)-G(x-z_1;\sigma)G(x-z_2;\sigma)]$$

$$[d\alpha_\tau(z_1,z_2)-d\alpha(z_1)d\alpha(z_2)]$$

$$+\int_{z_1}\int_{z_2}G(x-z_1;\sigma)G(x-z_2;\sigma)\;[d\alpha_\tau(z_1,z_2)-d\alpha(z_1)d\alpha(z_2)]. \tag{2.2.36}$$

We have added and subtracted the quantity

$$G(x-z_1;\sigma)G(x-z_2;\sigma)=\int_{-\infty}^{x}g(y-z_1;\sigma)dy\int_{-\infty}^{x}g(y-z_2;\sigma)dy$$

in the integrand.

Using our previous results (see (13)-(15)), we easily dominate the first expression on the right side of (36) by $(2|\rho_\tau|/(1-|\rho_\tau|))$. The bound for the variance becomes:

$$V(F_n(x))\le\frac{F(x)}{n}\,(1-F(x))$$

$$+\frac{6}{n^2}\sum_{\tau=1}^{n}(n-\tau)\,\frac{|\rho_\tau|}{1-|\rho_\tau|}$$

$$+\frac{2}{n^2}\sum_{\tau=1}^{n}(n-\tau)\int_{z_1}\int_{z_2}G(x-z_1;\sigma)G(x-z_2;\sigma)\,[d\alpha_\tau(z_1,z_2)-$$
$$d\alpha(z_1)d\alpha(z_2)]. \qquad (2.2.37)$$

Aside from the constant 6, the second expression is the same as (16). Hence, if $R(t) \to 0$ as $t \to \infty$, this term tends to zero. We now show that the ergodic condition on the $\{Z_\ell\}$ sequence is sufficient to have the third expression of (37) go to zero.

For the stationary sequence $\{Z_n\}$, $n = 0, \pm 1, \pm 2, \ldots$, a condition equivalent to ergodicity (Rosenblatt [34], p. 110) is:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}P(B\cap T^{-j}A) = P(A)P(B)\,, \qquad (2.2.38)$$

where A and B are any two events defined on the underlying probability space. P(A) denotes the probability of the set A and T is the unit shift transformation.

We take the elementary points of the probability space as $\omega = (\ldots,\omega_{-1},\omega_0,\omega_1,\ldots)$, where the ω_i are real numbers and define the random variable $Z_n(\omega) = \omega_n$. Equation (38) holds for any measurable set defined on the probability space. In particular, with $A = \{\omega | Z_{n_1}(\omega) \le z_1\}$, $T^{-j}A$ is the set

$$T^{-j}A = \{\omega | Z_{n_1}(T^j\omega) \le z_1\}$$
$$= \{\omega | Z_{n_1+j}(\omega) \le z_1\}\,.$$

Let $B = \{\omega | Z_{n_2}(\omega) \le z_2\}$. The stationarity assumption gives:

$$P(A) = \alpha_{Z_{n_1}}(z_1) = \alpha(z_1)$$
$$P(B) = \alpha_{Z_{n_2}}(z_2) = \alpha(z_2)$$

$$P(B \cap T^{-j}A) = \Pr(Z_{n_2}(\omega) \leq z_2,\ Z_{n_1+j}(\omega) \leq z_1) = \alpha_{n_1+j-n_2}(z_1,z_2) \ .$$

Define a second-order distribution function as

$$\beta_{n,n_1-n_2}(z_1,z_2) = \frac{1}{n} \sum_{j=1}^{n} \alpha_{n_1-n_2+j}(z_1,z_2) \ . \tag{2.2.39}$$

Then, equation (38) implies

$$\lim_{n\to\infty} \beta_{n,n_1-n_2}(z_1,z_2) = \alpha(z_1)\ \alpha(z_2) \tag{2.2.40}$$

for every z_1 and z_2 and every n_1 and n_2.

Returning to the variance expression, designate the double integration in (37) by $\tilde{D}_{\tau,2}$:

$$\tilde{D}_{\tau,2} = \int_{z_1}\int_{z_2} G(x-z_1;\sigma)\ G(x-z_2;\sigma)\ [d\alpha_\tau(z_1,z_2) - d\alpha(z_1)d\alpha(z_2)] \ . \tag{2.2.41}$$

Define the partial sum s_n by

$$s_n = \sum_{\tau=1}^{n} \tilde{D}_{\tau,2}$$

and the partial Cesaro sum S_n as

$$S_n = \frac{1}{n} \sum_{j=1}^{n} s_j$$

The third expression of (37) can then be written as

$$\frac{2}{n^2} \sum_{\tau=1}^{n} (n-\tau)\ \tilde{D}_{\tau,2} = \frac{2(n-1)}{n^2}\ S_{n-1} \ . \tag{2.2.42}$$

Consider the arithmetic mean of the partial sums,

$$\frac{s_n}{n} = \int_{z_1}\int_{z_2} G(x-z_1;\tau)G(x-z_2;\tau) \frac{1}{n}\sum_{\tau=1}^{n} [d\alpha_\tau(z_1,z_2) - d\alpha(z_1)d\alpha(z_2)] \tag{2.2.43}$$

From the ergodic hypothesis and equations (38)-(40), as $n \to \infty$, we have

$$\frac{1}{n}\sum_{\tau=1}^{n} \alpha_\tau(z_1,z_2) \longrightarrow \alpha(z_1)\ \alpha(z_2)$$

for every z_1 and z_2. Applying the Helly-Bray Theorem (Loève, [24], p. 183), we get that $s_n/n \to 0$ as $n \to \infty$. It then follows (see the discussion below (19)) that the arithmetic mean of the partial Cesaro sums $S_{n-1}/n \to 0$. Hence, (42) and $V(F_n(x))$ tend toward zero as $n \to \infty$.

We have just shown that an ergodic assumption on the sequence Z and a somewhat stronger assumption on the gaussian noise implies that $V(F_n(x)) \to 0$.[1] Also, we are now in a position to investigate the rate of convergence.

If we were dealing with independent samples, the rate of convergence would be $O(1/n)$. For definiteness, we will consider this particular rate.

Clearly, if $R(t)$ satisfies condition B, the second expression on the right side of (37) will be $O(1/n)$. For the third expression of (37) to be of the same order, the sequence of Cesaro sums, S_n, must either oscillate between finite bounds or converge. There are necessary and sufficient conditions for Cesaro summability. For example, we have:[2]

---

[1] We will discuss this assumption on the gaussian noise later.

[2] Knopp, K., Theory and Application of Infinite Series, Hafner Pub. Co., 1950, p. 486. Translated from the second German Edition.

a necessary and sufficient condition for a series $\sum a_n$, with partial sums s_n, to be Cesaro summable to the sum S is that the series

$$\text{i)} \quad \sum_{v=0}^{\infty} \frac{a_v}{v+1}$$

should be convergent and that for its remainder

$$\rho_n = \frac{a_{n+1}}{n+2} + \frac{a_{n+2}}{n+3} + \ldots \quad (n = 0,1, \ldots)$$

the relation

$$\text{ii)} \quad S_n + (n+1)\rho_n \longrightarrow S \text{ holds.}$$

With $\tilde{D}_{\tau,2} = a_\tau$, the above conditions, in conjunction with condition B on R(t), yield the rate 1/n. These conditions, however, are difficult to interpret in terms of the $\{Z_\ell\}$ process. We shall content ourselves with a simple sufficient condition which admits some interpretation and which at the same time is not an overly restrictive assumption for the type of problems we will want to deal with.

The Cesaro method is regular; that is, if $\sum_\tau \tilde{D}_{\tau,2}$ converges to s, then S_n also converges to s. Hence, a sufficient condition to achieve the rate 1/n is:

$$\sum_{\tau=1}^{\infty} \tilde{D}_{\tau,2} = B_3 < \infty. \qquad (2.2.44)$$

As our final result, we have

Theorem 2.2.2: Given the sequence of identically distributed random variables, $X_i = N_i + Z_i$, $i=1,2,\ldots$. Assume:

i) $R(t)$ satisfies condition B

ii) $\tilde{D}_{\tau,2}$ as given by (41) satisfies (44).

Then, the empirical distribution function is a consistent estimate of $F(x)$ with the variance dominated by

$$E(F_n(x) - F(x))^2 = V(F_n(x)) \leq \frac{1}{n}\left(1 + \frac{6\,B_2}{1-\rho^*} + B_3\right) .$$

In addition, if

iii) the sequence $\{Z_\ell\}$ is ergodic, $F_n(x)$ converges to $F(x)$ with probability one for every x.

The term $2B_2$ is contributed by the first part of $D_{m-\ell}$ which is due to the correlated noise, while $4B_2$ results from the bound on the first part of $D_{\tau,2}$.

Among other things, (44) implies that $\tilde{D}_{\tau,2} \to C$ as $\tau \to \infty$. We have seen that ergodicity implies only $\frac{1}{n}\sum_{\tau=1}^{n} \tilde{D}_{\tau,2} \to C$ as $n \to \infty$. To characterize the type of process which satisfies the assumptions of the theorem, we can replace the ergodic assumption on the $\{Z_\ell\}$ (equation (38)) by the stronger condition

$$\lim_{n\to\infty} P(B \cap T^{-j}A) = P(A)P(B) . \qquad (2.2.45)$$

This is called a mixing condition (Rosenblatt [34], p. 110). The mixing condition implies that $\alpha_\tau(z_1,z_2) \to \alpha(z_1)\,\alpha(z_2)$ for every z_1 and z_2, and from the Helly-Bray Theorem we have $\lim_{\tau\to\infty} \tilde{D}_{\tau,2} = 0$. Thus, one class of

processes which satisfies ii) of the above theorem is the mixing processes whose dependency is weak enough so that

$$\sum_{\tau=1}^{\infty} \int_{z_1} \int_{z_2} |d\alpha_\tau(z_1,z_2) - d\alpha(z_1)d\alpha(z_2)| < \infty \qquad (2.2.46)$$

is satisfied. From the definition of $\tilde{D}_{\tau,2}$, this condition then implies ii) of the above theorem.

We observe that if the sequence $\{Z_\ell\}$ is M-dependent, it satisfies the mixing condition.[1] In this case, ii) of the theorem and (46) are obviously satisfied. If the $\{Z_\ell\}$ sequence were gaussian, condition ii) would be satisfied if its autocorrelation function satisfied condition B.

A further characterization of processes which satisfy condition ii) is given in section 2.7.

The condition $R(t) \to C$ as $t \to \infty$, as we have already remarked, implies ergodicity. The implication does not go the other way. In fact, this condition on the correlation function is both a necessary and sufficient condition for the gaussian process to be mixing.[2]

What we have done in this section is to employ the Mehler formula to dominate the integral

---

[1] Rosenblatt, [34], p. 110, shows that a stationary process of independent, identically distributed random variables satisfies the mixing condition. The extension to a M-dependent process is easy.

[2] Rosenblatt, M., "Independence and Dependence," Proc. 4th Berkeley Symposium Math. Statistics and Probability (1961) v. 2, pp. 431-443.

$$\iint |g_2(x_1-z_1, x_2-z_2; \sigma, \rho_\tau) - g(x_1-z_1;\sigma)g(x_2-z_2;\sigma)| \, dx_1 \, dx_2 \, .$$

The bound we have obtained is independent of z_1 and z_2 and is given in terms of the correlation coefficient ρ_τ. By specifying the manner in which the correlation function $R(\tau)$ goes to zero, we then obtained a bound on $V(F_n(x))$.

When the Z_i are dependent, we have to require a condition like (44) so as to specify a rate of convergence.

In estimating the density function $f(x)$, we will make use of these results. In all three methods which we present the variance of the estimate is dominated in the same manner; the part of the variance expression which is due to the dependency of the observations is written as the difference of two expectations involving the appropriate bivariate and univariate density functions.

2.3 ESTIMATE OF THE DENSITY FUNCTION—KERNEL METHOD

In this section we consider a method of estimating the density function which is analogous to that used in estimating the spectral density of a stationary time series. This approach has already been applied to the case of a sequence of independent random variables [25,26,33,46]. We will generally follow Parzen [26].

The density function we want to estimate is given in (2.1.4) and repeated here,

$$f(x) = \int_z g(x-z;\sigma) d\alpha(z) \, . \qquad (2.3.1)$$

From the observations X_i, i=1,2,...n, we take an estimate of the form:

$$\hat{f}_n(X_1,X_2,\ldots X_n;x) = \hat{f}_n(x) = \frac{1}{nh(n)} \sum_{\ell=1}^{n} K\left(\frac{x-X_\ell}{h(n)}\right) \qquad (2.3.2)$$

where h is a sequence of positive numbers depending on n, and chosen so that

$$\lim_{n\to\infty} h(n) = 0 . \qquad (2.3.3)$$

K(x) is a non-negative function satisfying

$$\begin{aligned} &\sup_{-\infty < x < \infty} K(x) < \infty \\ &\int K(x)dx < \infty \\ &\lim_{x\to\infty} |xK(x)| = 0 \end{aligned} \qquad (2.3.4)$$

2.3a Bias Calculation

The expectation of (2) is

$$E\,\hat{f}_n(x) = \frac{1}{h} E\left\{K\left(\frac{x-X}{h}\right)\right\} = \frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)dy \qquad (2.3.5)$$

The following theorem (specialized to our situation) is given in Parzen.

Theorem 2.3.1 (Parzen, p. 1067): With h and K(y) satisfying (3), (4), respectively, we have

$$\lim_{n\to\infty} E\, f_n(x) = f(x)\int_{-\infty}^{+\infty} K(y)dy \qquad (2.3.6)$$

at every point x of continuity of f( ).

With $\int K(y)dy = 1$, and since $f(x)$ is everywhere continuous, $\hat{f}_n(x)$ is asymptotically unbiased estimate of $f(x)$ for every x. In fact, since the gaussian density is uniformly continuous, $f(x)$ is also uniformly continuous. It then follows from Parzen's proof that the convergence is uniform in x.

Our particular density $f(x)$ is more specialized than that needed for the proof of the theorem. We can use some of its properties to obtain a uniform bound on the bias. Rewrite (5) as

$$E\,\hat{f}_n(x) = \int_{-\infty}^{+\infty} K(u)f(x-hu)du \ . \tag{2.3.7}$$

Since $\int K(u)du = 1$, we can write

$$E\,\hat{f}_n(x) - f(x) = \int_{-\infty}^{+\infty} K(u)[f(x-hu) - f(x)]du \tag{2.3.8}$$

To find the limiting behavior of the integral, we expand $f(x-hu)$ in a Taylor series about the point x. Since $f(x)$ is the convolution of a distribution with a gaussian density, all derivatives of $f(x)$ exist.

$$f(x-hu) = f(x) - hu\,f'(x) + \frac{h^2u^2}{2} f''(x) + O(h^3) \tag{2.3.9}$$

Choose $K(u)$ as an even function and require that

$$\int_{-\infty}^{+\infty} u^2\,K(u)du = B_4 < \infty \ . \tag{2.3.10}$$

Two examples of even, non-negative kernels which satisfy this condition (as well as (4)) are:

$$K(u) = \frac{1}{2} \quad , \quad |u| \leq 1$$

$$= 0 \quad , \text{ otherwise}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} \, e^{-\frac{u^2}{2}} \quad .$$

They also integrate to one.

Substitute the Taylor series into (8), and perform the integrations. As $n \to \infty$, we get

$$E \, \hat{f}_n(x) - f(x) \longrightarrow \frac{f''(x)}{2} \, B_4 \, h^2 \quad . \tag{2.3.11}$$

To obtain a uniform bound (and for future reference) we note that the derivatives of $f(x)$ are uniformly bounded in x. Specifically, from (A.9) it follows that

$$\frac{d^j f(x)}{dx^j} = \int_z \frac{d^j}{dx^j} g(x-z;\sigma) d\alpha(z) = \frac{(-1)^j}{\sigma_j} \int_z He_j \left(\frac{x-z}{\sigma}\right) g\left(\frac{x-z}{\sigma}\right) d\alpha(z)$$

and

$$\left| \frac{d^j f(x)}{dx^j} \right| \leq \frac{1}{\sigma^{j+1}} \frac{c_1 \sqrt{j!}}{\sqrt{2\pi}} \int_z d\alpha(z) = \frac{c_1}{\sqrt{2\pi}} \frac{\sqrt{j!}}{\sigma^{j+1}} \quad . \tag{2.3.12}$$

The last line follows from Cramer's bound, (A.30). With $j = 2$, we have

$$|f''(x)| \leq c_1 / \sqrt{\pi} \, \sigma^3$$

and as $n \to \infty$ (11) is dominated by

$$|E \, \hat{f}_n(x) - f(x)| \leq \frac{c_1 \, B_4 \, h^2}{2\sqrt{\pi} \, \sigma^3} + O(h^4) \quad . \tag{2.3.13}$$

It is advantageous, in terms of bias to have h(n) go to zero rapidly. Consideration of the variance of the error, however, will show that it should not approach zero too rapidly.

2.3b Variance Calculation

The square of $\hat{f}_n(x)$ is written as

$$\hat{f}_n^{\,2}(x) = \frac{1}{n^2h^2} \sum_{\ell=1}^{n} K^2\left(\frac{x-X_\ell}{h}\right) + \frac{2}{n^2h^2} \sum_{\ell=1}^{n} \sum_{m=\ell+1}^{n} K\left(\frac{x-X_\ell}{h}\right) K\left(\frac{x-X_m}{h}\right) \qquad (2.3.14)$$

We proceed in a manner analogous to the development in section 2.2. Take the expectation of (14), subtract

$$\frac{2}{n^2h^2} \sum_{\ell=1}^{n} \sum_{m=\ell+1}^{n} E\left\{K\left(\frac{x-X_\ell}{h}\right)\right\} E\left\{K\left(\frac{x-X_m}{h}\right)\right\}$$

and add its equivalent

$$\frac{1}{h^2}\left(1 - \frac{1}{n}\right) \left[E\left\{K\left(\frac{x-X}{h}\right)\right\}\right]^2 .$$

Subtracting the square of the bias, we obtain the variance:

$$\begin{aligned} V(\hat{f}_n(x)) &= E(\hat{f}_n^{\,2}(x) - E\,\hat{f}_n(x))^2 \\ &= \frac{1}{nh^2}\left[E\left\{K^2\left(\frac{x-X}{h}\right)\right\} - \left[E\left\{K\left(\frac{x-X}{h}\right)\right\}\right]^2\right] \\ &+ \frac{2}{n^2h^2} \sum_{\ell=1}^{n} \sum_{m=\ell+1}^{n} \left[E\left\{K\left(\frac{x-X_\ell}{h}\right)K\left(\frac{x-X_m}{h}\right)\right\} - E\left\{K\left(\frac{x-X_\ell}{h}\right)\right\} E\left\{K\left(\frac{x-X_m}{h}\right)\right\}\right] . \end{aligned} \qquad (2.3.15)$$

Again, the second term is a result of the dependency of the observations. We proceed to majorize each of the terms in the variance expression.

The first term written out is

$$\frac{1}{nh^2}\ E\left\{K^2\left(\frac{x-X}{h}\right)\right\} = \frac{1}{nh^2}\int_{-\infty}^{+\infty} K^2\left(\frac{x-y}{h}\right) f(y)dy\ . \qquad (2.3.16)$$

From the first two conditions of (4) we have $\int K^2(y)dy = B_5 < \infty$. Since $f(y) \leq 1/\sqrt{2\pi}\sigma$, the substitution of $u = (x-y)/h$ in (16) leads to

$$\frac{1}{nh^2}\ E\left\{K^2\left(\frac{x-X}{h}\right)\right\} \leq \frac{B_5}{nh\sqrt{2\pi}\ \sigma} \qquad (2.3.17)$$

The second term of (15) is

$$\frac{1}{nh^2}\left[E\left\{K\left(\frac{x-X}{h}\right)\right\}\right]^2 = \frac{1}{nh^2}\left\{\int_{-\infty}^{+\infty} K\left(\frac{x-y}{h}\right)\ f(y)dy\right\}^2\ . \qquad (2.3.18)$$

Use the boundedness of $f(y)$, the same substitution as above, and the fact that $K(z)$ is non-negative and intergrates to one to obtain

$$\frac{1}{nh^2}\left[E\left\{K\left(\frac{x-X}{h}\right)\right\}\right]^2 \leq \frac{1}{n}\left(\frac{1}{2\pi\sigma^2}\right)\ . \qquad (2.3.19)$$

For the third term of the variance equation, let $Q_{m-\ell}$ be the expression inside the double sum. Writing this term out gives

$$Q_{m-\ell} = \iint K\left(\frac{x-y_1}{h}\right)K\left(\frac{x-y_2}{h}\right)\left[f_{m-\ell}(y_1,y_2)-f(y_1)f(y_2)\right]dy_1dy_2\ , \qquad (2.3.20)$$

where $f_{m-\ell}(y_1,y_2)$ is the second-order density function of the observations in the m and ℓ intervals. The kernel $K(y)$ is bounded (4), say by B_7. Hence,

$$Q_{m-\ell} \leq B_7^2 \iint |f_{m-\ell}(y_1,y_2)-f(y_1)f(y_2)|dy_1dy_2\ , \qquad (2.3.21)$$

and we are in a position to use the results of the previous section (see equations (2.2.10)-(2.2.15)). For example, with the sequence $\{Z_\ell\}$ independent,

$$Q_{m-\ell} \leq B_7^2 \frac{|\rho_{m-\ell}|}{1 - |\rho_{m-\ell}|} \tag{2.3.22}$$

Combining this with (17) and (19), (15) is majorized by

$$V(\hat{f}_n(x)) \leq \frac{1}{n} \frac{1}{2\pi\sigma^2} + \frac{1}{nh} \frac{B_5}{\sqrt{2\pi}} + \frac{2B_7^2}{h^2n^2} \sum_{\ell=1}^{n} \sum_{m=\ell+1} \frac{|\rho_{m-\ell}|}{1 - |\rho_{m-\ell}|} .$$

With the autocorrelation function satisfying condition B, (2.2.23) gives

$$V(\hat{f}_n(x)) \leq \frac{1}{n} \frac{1}{2\pi\sigma^2} + \frac{1}{nh} \frac{B_5}{\sqrt{2\pi}\,\sigma} + \frac{2B_7^2 B_2}{nh^2(1-\rho_*)} . \tag{2.3.23}$$

Under these conditions it follows that we need to require $nh^2 \to \infty$ as $n \to \infty$ for consistency of the estimate. Notice that if the observations were independent, the third term would be absent. In this case (as in Parzen's development) we need only require $nh \to \infty$.

In Appendix C, section C.1, we show that by choosing K(u) as the gaussian kernel,

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} ,$$

(22) is replaced by

$$Q_{m-\ell} \leq h^2 \frac{|\rho_{m-\ell}|}{1 - |\rho_{m-\ell}|} \frac{c_1^2}{2\pi\sigma} . \tag{2.3.24}$$

The reason this is possible is that with the specific gaussian kernel we can perform the y integrations in (20) before taking bounds. In this case, the variance is then dominated by

$$V(\hat{f}_n(x)) \leq \frac{b_1}{n} + \frac{b_2}{nh} , \qquad (2.3.25)$$

where we have set

$$b_1 = \frac{1}{2\pi\sigma^2} + \frac{2B_2}{(1-\rho_*)} \cdot \frac{c_1}{2\pi\sigma}$$

$$b_2 = B_5\sqrt{2\pi} .$$

The rest of the discussion will assume a gaussian kernel, and for definiteness, we assume that the autocorrelation function satisfies condition B. It was this condition on the noise that gave, for the empirical distribution function, a rate of convergence equal to the case of independent observations. We expect analogous results for estimating the density function.

2.3c Mean-Square Error

The mean-square error is written in terms of the bias and variance contributions.

$$E\left[\hat{f}_n(x) - f(x)\right]^2 = E(\hat{f}_n(x) - E(\hat{f}_n(x))^2 + \left[E(\hat{f}_n(x) - f(x))\right]^2$$

As $n \to \infty$, from (13) and (25), and setting $\sqrt{b_3} = (c_1B_4)/(\sqrt{\pi}\, 2\sigma^3)$, we have

$$E\{f_n(x) - f(x)\}^2 \leq \frac{b_1}{n} + \frac{b_2}{nh} + b_3 h^4 . \qquad (2.3.26)$$

Clearly, to minimize this bound, we choose h (as a function of n) so as to have the last two terms of (26) tend to zero at the same rate. Differentiating, we find that the best h is given by

$$h = \left[\frac{b_2}{4b_3 n}\right]^{1/5} \qquad (2.3.27)$$

Therefore, as $n \to \infty$, the mean-square error satisfies

$$E(\hat{f}_n(x) - f(x))^2 = O(1/n^{4/5}). \qquad (2.3.28)$$

This is the order of consistency one obtains for the case of independent samples [26,33].

Theorem 2.3.2: The estimate of the form

$$f(X_1, X_2, \ldots X_n, X) = \frac{1}{n} \sum_{\ell=1}^{n} g(x - X_\ell; h(n))$$

converges in mean-square, uniformly in x, at a rate $1/n^{4/5}$ if:

i) $\{Z_\ell\}$ are independent

ii) $R(\tau)$ satisfies condition B

iii) $h(n)$ is chosen as in (2.3.27)

Clearly, we can extend the results for other dependencies on Z. For the M-dependent case, we have:

Corollary 2.3.1: Under the preceeding hypotheses and with condition i) replaced by

i) Z_ℓ is independent of Z_m if $|m-1| \geq M$, the order of consistency remains = $O(1/n^{4/5})$

This corollary follows from the comments in appendix C, (C.10), choosing h as in (27), and taking b_1, as

$$b_1 = \frac{1}{2\pi\sigma^2}\left[1 + \frac{4(M-1)}{1-\rho_*}\right] . \qquad (2.3.29)$$

We have obtained the result that the estimate $\hat{f}_n(X_1,\ldots X_n;x)$ converges in mean-square to the univariate density at a rate not slower than $1/n^{4/5}$. As will be seen in the next chapter, in attempting to estimate the k-variate density function $f(x_1,\ldots x_k)$, the bound on convergence which we are able to specify indicates slower convergence.

Another disadvantage of this method is the problem of "growing memory." The estimate we have been using is of the form

$$\hat{f}_n(X_1,X_2,\ldots X_n;x) = \frac{1}{nh(n)} \sum_{\ell=1}^{n} K\left(\frac{x-X_\ell}{h(n)}\right) ,$$

from which it can be seen that all past observations must be stored—at each stage a particular observation's contribution to the estimate is weighted differently.

This problem can be eliminated if one is willing to accept a final estimate which is biased. For this situation, we need only store the past N observations, where N is determined by tne bias one will accept. A recursive relationship is then used to update the estimate.

One advantage of the kernel method is that the estimate is a density function; $\hat{f}_n(x)$ in non-negative and integrates to one. Another advantage is that no knowledge of the gaussian or z process is required to form

the estimate, and only a minimal amount of information is required to specify the rate of convergence.

2.3d Mean Integrated Square Error (MISE)

Another criterion which has been used to measure the error is the mean integrated square error (MISE). Using this criterion, one can specify an optimum choice of the kernel $K(y)$ and investigate maximum rates of convergence. It is primarily the rate of convergence which we now want to discuss.

The MISE is defined as

$$J_n = E\left[\int (\hat{f}_n(X_1,\ldots,X_n;x) - f(x))^2\, dx\right] \qquad (2.3.30)$$

We remark that the condition $nh(n) \to \infty$ is sufficient to show that, with probability one,

$$f_n(X_1,\ldots X_n;x) = \frac{1}{n}\sum_{\ell=1}^{n} g(x-X_\ell;h(n)) \qquad (2.3.31)$$

is square integrable in x. In fact, in appendix C we obtain a bound on the MISE. Our result (see section C.2) is that the MISE = $O(1/n^{4/5})$, which is the same rate obtained for the mean-square error.[1] The question naturally arises as to whether one can specify a maximum rate of convergence using estimators of the above type.

[1] This is for the case of independent Z and correlated noise with condition B holding.

Watson and Leadbetter [45] consider the problem of optimizing the estimator of the form

$$\hat{f}_n(x) = \frac{1}{n} \sum_{\ell=1}^{n} K_n(x-X_\ell) \qquad (2.3.32)$$

given a sequence of independent observations. They show that the MISE is a minimum if the Fourier transform of the kernel $K_n(x)$, which we denote by $M_{K_n}(v)$, is equal to

$$M_{K_n}(v) = \frac{|M_f(v)|^2}{\frac{1}{n} + \frac{(n-1)}{n} |M_f(v)|^2} \quad . \qquad (2.3.33)$$

$M_f(v)$ is the characteristic function of $f(x)$. Notice that $M_{K_n}(v) \to 1$ as $n \to \infty$, indicating that $K_n(x)$ approaches a delta function as is the case with the previous estimator. Here, however, the kernel's functional form is dependent on the index n, which gives a more complicated estimator.

They show that the minimum MISE cannot decrease faster than $1/n$. Specifically, with the optimum kernel given by the inverse transform of (33), the minimum MISE is

$$J_{n*} = \frac{M_{K_n}(0)}{n} - O(1/n)$$

Watson and Leadbetter further characterize the optimum estimator by studying the asymptotic behavior of the (unknown) characteristic function $M_f(v)$.[1]

[1]The estimator in (33) is of no practical value since it is expressed in terms of the function being estimated. By specifying the asymptotic behavior of $M_f(v)$, they show that there is a class of kernels, with the same asymptotic behavior, which achieves the maximum rate of convergence.

Of particular interest to us is the class of characteristic functions which decrease exponentially with degree r and coefficient γ. A characteristic function is of this class if it satisfies:

i) $|M_f(v)| \leq A\, e^{-\gamma|v|^r}$, for some constants $A > 0$, $\gamma > 0$ and $0 \leq r \leq 2$.

ii) $$\int_0^1 \frac{dt}{1 + \exp(2\gamma v^r)\, |M_f(tv)|^2} \to 0 \text{ as } v \to \infty .$$

Under these conditions they state the following theorem.

Theorem 2.3.3: (Watson and Leadbetter, p. 490): Let $M_f(v)$ decrease exponentially with coefficient γ and degree r. Then J_{n*}, the minimum MISE, satisfies

$$\lim_{n\to\infty} [n/(\log n)^{1/r}]\, J_{n*} = [1/\pi(2\gamma)^{1/r}]$$

For our case r=2 and with independent observations, the minimum MISE converges to zero at a rate $\sqrt{\log(n)/n}$. This assumes an estimate with the kernel given by (33) and represents an improvement in the convergence rate at the expense of a more complicated estimator.

With the noise correlated and the sequence $\{Z_\ell\}$ independent, we can obtain a corresponding expression for the optimum kernal expressed in terms of $M_f(v)$ and the sequence of correlation coefficients $\{\rho_\tau\}$. Calculating the rate of decrease for the MISE is difficult. However, it is easy to show that the minimum MISE cannot decrease faster than $1/n$.[1]

[1] These comments are substantiated by paralleling the development in [45].

It is this rate and the one for the independent case which we will want to use as a point of reference while discussing the series methods of estimating f(x).

2.4 ESTIMATING THE DENSITY FUNCTION BY SERIES METHODS—AN ORTHOGONAL REPRESENTATION

In this section f(x) is represented in an orthogonal expansion and it is this form which we will estimate.[1] Our concern will be with convergence, not only in the sense of MISE, but in mean-square as given by (2.1.1) and (2.1.2).

The density function

$$f(x) = \int g(x-z;\sigma)d\alpha(z)$$

is a bounded integrable function. Hence, it is L_2 and can be expanded in a series of orthonormal functions:

$$f(x) = \sum_{j=0}^{\infty} a_j \varphi_j(x/\sigma_1) \qquad (2.4.1)$$

$$a_j = \int_{-\infty}^{+\infty} f(x)\, \varphi_j(x/\sigma_1)dx \ . \qquad (2.4.2)$$

Naturally, equality in (1) is in the sense of limit-in-mean. The functions in the expansion are the normalized Hermite functions which form a complete orthonormal set on the whole line (see Appendix A, section A.2):

[1] This technique has been discussed in [6], but not in any depth.

$$\varphi_j(x/\sigma_1) = \frac{g(x/\sigma_1)\ H_j(x/\sigma_1)}{\sqrt{2^j j!}/\sqrt{4\pi}\ \sigma_1} \tag{2.4.3}$$

σ_1 is an arbitrary positive constant.

As discussed in appendix A, we reserve the H notation for the polynomials orthogonal with respect to the square of the gaussian weight. They are generated by differentiating $g^2(x)$. The He polynomials are generated by differentiating $g(x)$ and are orthogonal with respect to the gaussian weight. These polynomials were introduced earlier.

Given the sequence of observations $\{X_\ell\}$, $\ell=1,2,\ldots,n$, the problem of estimating $f(x)$ (in the L_2 sense) is reduced to one of estimating the coefficients a_j. We designate the estimate of the a_j coefficient at the n-th state by $\hat{a}_{j_n}$:

$$\hat{a}_{j_n} = \frac{1}{n} \sum_{\ell=1}^{n} \varphi_j(X_\ell/\sigma_1). \tag{2.4.4}$$

It follows that these estimates are unbiased:

$$E\ \hat{a}_{j_n} = \frac{1}{n} \sum_{\ell=1}^{n} E\ \varphi_j(X_\ell/\sigma_1)$$

$$= \int_{-\infty}^{+\infty} \varphi_j(x/\sigma_1)\ f(x)dx = a_j\ . \tag{2.4.5}$$

The mean-square error in the estimate is then given by the variance of $\hat{a}_{j_n}$. To calculate this variance we proceed as before:

$$V(\hat{a}_{j_n}) = E(\hat{a}_{j_n} - E\ \hat{a}_{j_n})^2 = E(\hat{a}_{j_n})^2 - a_j^2$$

$$= E\left\{\frac{1}{n^2}\sum_{\ell=1}^{n}\varphi_j^2(X_\ell/\sigma_1) + \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}\varphi_j(X_\ell/\sigma_1)\varphi_j(X_m/\sigma_1)\right\} - a_j^2$$

$$= \frac{1}{n}\left\{E(\varphi_j^2(X/\sigma_1)) - \left[E(\varphi_j(X/\sigma_1))\right]^2\right\}$$

$$+ \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}\left\{E(\varphi_j(X_\ell/\sigma_1)\,\varphi_j(X_m/\sigma_1)) - E(\varphi_j(X_\ell/\sigma_1))\right.$$

$$\left. E\,(\varphi_j(X_m/\sigma_1))\right\} . \qquad (2.4.6)$$

The first expression on the right is easily dominated using Cramer's bound. From (A.29), we have

$$e^{-\frac{x}{2}^2}\frac{|H_j(x)|}{\sqrt{2^j j!}} < c_1 , \qquad (A.29)$$

where the constant c_1 is independent of x and j. Hence,

$$|\varphi_j(x/\sigma_1)| < c_1/(\pi^{1/4}\,\sigma_1^{1/2}) = c_2 , \qquad (2.4.7)$$

and the first expression in (6) is dominated by $2c_2^2/n$.

For the second expression in (6), we again write out the expectations and use the above bound on $\varphi_j(x)$ to dominate it by

$$\frac{2}{n^2}\,c_2^2\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|f_{m-\ell}(x_1,x_2) - f(x_1)f(x_2)|dx_1dx_2 . \qquad (2.4.8)$$

We have already majorized this term. As it will again appear in the sequel, we use our previous results to record the following lemmas.

Lemma 2.4.1: Assume

i) the autocorrelation function satisfies condition B

ii) the sequence $\{Z_\ell\}$ is M-dependent.

Then, (8) is majorized by

$$\frac{c_2^2}{n}\left(\frac{2B_2}{1-\rho^*} + 4(M-1)\right)$$

Alternatively, if $|R(\tau)| \leq \sigma^2/\tau^\delta$ for $|\tau| > B_1$ and $0 < \delta < 1$ (condition A), we have

Lemma 2.4.2: Assume

i) $R(\tau)$ satisfies condition A

ii) the sequence $\{Z_\ell\}$ is M-dependent

Then, (8) is majorized by

$$\frac{2c_2^2}{1-\rho_*}\left(\frac{B_1}{n} + \frac{1}{(1-\delta)n^\delta}\right) + \frac{4c_2^2(M-1)}{n} \quad . \tag{2.4.9}$$

The proofs and appropriate definitions are given in the discussion leading to Corollary 2.2.1. Clearly, the M-dependence assumption can be replaced by a weaker condition as reflected in Theorem 2.2.2. Again, the point here is that the estimate of a_j is taken as an average of bounded functions and the problem then reduces to one of dominating a sum involving the absolute difference of the bivariate and univariate density functions.

In the following discussion we shall refer only to Lemma 2.4.1.

Theorem 2.4.1: The estimate as given by

$$\hat{a}_{jn} = \frac{1}{n}\sum_{\ell=1}^{n} \varphi_j(X_\ell/\sigma_1) \tag{2.4.4}$$

is unbiased, and under the hypotheses of Lemma 2.4.1, the variance is bounded by

$$V(\hat{a}_{jn}) \leq \frac{c_2^2}{n}\left(2 + \frac{2\,B_2}{1-\rho_*} + 4(M-1)\right) = \frac{c_3}{n} \quad . \tag{2.4.10}$$

Now define the estimate of the density function at the n-th stage by

$$\hat{f}_n(x) = \sum_{j=0}^{q(n)} \hat{a}_{jn}\, \varphi_j(x/\sigma_1) \ , \tag{2.4.11}$$

where $q(n)$ is an integer which depends on n. Consider the MISE:

$$\begin{aligned} J_n &= \int_{-\infty}^{+\infty} E(\hat{f}_n(x) - f(x))^2 \, dx \\ &= \sum_{j=q+1}^{\infty} a_j^2 + \sum_{j=1}^{q} E(\hat{a}_{jn} - a_j)^2 \\ &\leq \sum_{j=q(n)+1}^{\infty} a_j^2 + \frac{c_3\, q(n)}{n} \quad . \end{aligned} \tag{2.4.12}$$

Since $\sum_{j=1}^{\infty} a_j^2 = \int f^2(x)dx < \infty$, it follows that the first term of (12) goes to zero if $q(n) \to \infty$. With $q(n)$ properly chosen so that the ratio $q(n)/n \to 0$, we have $J_n \to 0$.

The problem of specifying the sequence $\{q(n)\}$, $n=1,2,\ldots$, is analogous to choosing the constants $h(n)$ in the previous section so as to balance the bias and variance errors. Here, a partial answer is provided if we assume that the random variable Z has, for example, a finite second moment,

$\int_{-\infty}^{+\infty} z^2 d\alpha(z) < \infty$. Below, we will show that this assumption implies that the coefficients a_j satisfy

$$a_j^2 \leq B_\theta / j^2 , \tag{2.4.13}$$

where B_θ is a constant independent of j. With this being the case, from the Euler-Maclaurin summation formula, we obtain:

$$\sum_{q(n)+1}^{\infty} a_j^2 < B_\theta \sum_{q+1}^{\infty} \frac{1}{j^2} \doteq B_\theta \int_q^{\infty} (\frac{1}{t^2} + \frac{1}{t^3}) dt = B_\theta (\frac{1}{q} + \frac{1}{q^2}). \tag{2.4.14}$$

The MISE is then dominated by

$$J_n < B_\theta (\frac{1}{q} + \frac{1}{q^2}) + \frac{c_\theta q(n)}{n}$$

and with $q(n) = n^{1/2}$,[1]

$$J_n = O(1/n^{1/2}). \tag{2.4.15}$$

This does not compare favorably with the $1/n^{4/5}$ rate obtained by the kernel method. Here, however, the rate $1/n^{1/2}$ is not a direct result of the method of estimation, but rather, results from the second moment assumption. After proving (13), it will be clear as to what additional moment assumption is needed to achieve any rate up to 1/n.

Lemma 2.4.3: With $\int_{-\infty}^{+\infty} z^2 d\alpha(z) < \infty$, the coefficients a_j satisfy

---

[1] More precisely, since q is an integer, we choose $q = [\sqrt{n}]$, where $[\ ]$ denotes the largest integer $\leq \sqrt{n}$.

$$a_j^2 \leq B_\theta / j^2 \ .$$

To prove this lemma we first obtain the conditions on $f(x)$ needed to give the above bound and then show that the second moment assumption implies these conditions.[1]

Proof: From (3),

$$a_j = \int_{-\infty}^{+\infty} f(x)\varphi_j(x/\sigma_1)dx = \int_{-\infty}^{+\infty} f(x) \frac{g(x/\sigma_1)H_j(x/\sigma_1)dx}{\sqrt{2^j j!\,/\sqrt{4\pi}\,\sigma_1}} \qquad (2.4.16)$$

We note the relationship for the derivative of the Hermite polynomial which is obtained from (A.28),

$$\frac{d}{dx} H_{j+1}(x/\sigma_1) = \frac{2(j+1)}{\sigma_1} H_j(x/\sigma_1) \ .$$

Substitute for $H_j(x/\sigma_1)$ and integrate by parts.

$$a_j = \int_{-\infty}^{+\infty} f(x) \frac{g(x/\sigma_1)}{\sqrt{2^j j!\,/\sqrt{4\pi}\,\sigma_1}} \frac{\sigma_1 \frac{d}{dx} H_{j+1}(x/\sigma_1)}{2(j+1)} dx$$

$$= \frac{-1}{\sqrt{2(j+1)}} \int_{-\infty}^{+\infty} \frac{H_{j+1}(x/\sigma_1)}{\sqrt{2^{j+1}(j+1)!\,/\sqrt{4\pi}\,\sigma_1}} \sigma_1^2 \left[\frac{d}{dx} f(x)g(x/\sigma_1)\right] dx =$$

$$= \frac{1}{\sqrt{2(j+1)}} \int_{-\infty}^{+\infty} \varphi_{j+1}(x/\sigma_1) \left[x f(x) - f'(x)\,\sigma_1^2\right] dx \ . \qquad (3.4.17)$$

Repeat the argument with the function $(xf(x) - \sigma_1^2 f'(x))$ playing the role

---

[1] The essential idea of the lemma can be found in Sansone [38], pp. 368-369.

of f(x):

$$a_j = \frac{-1}{\sqrt{4(j+1)(j+2)}} \int_{-\infty}^{+\infty} \frac{H_{j+2}(x/\sigma_1)\ \sigma_1^{\,2}\ \frac{d}{dx}\left\{g(x/\sigma_1)(xf(x)-\sigma_1^{\,2}f'(x))\right\}}{\sqrt{2^{j+2}\ (j+2)!/\sqrt{4\pi}\ \sigma_1}}\,dx$$

$$= \frac{1}{2\sqrt{(j+1)(j+2)}} \int_{-\infty}^{+\infty} \varphi_{j+2}(x/\sigma_1)\ w(x)\ dx, \qquad (2.4.18)$$

where $w(x) = ((x^2-\sigma_1^{\,2})f(x) - 2\sigma_1^{\,2}xf'(x) + \sigma_1^{\,4}f''(x))$. If $w(x)\epsilon L_2$ we obtain

$$a_j^{\,2} \leq \frac{1}{4(j+1)(j+2)} \int_{-\infty}^{+\infty} w^2(x)dx \leq \frac{B_6}{j^2} \qquad (2.4.19)$$

where B_6 is the L_2 norm of $w(x)/2$.

Using characteristic functions, it is easy to see that $w(x)$ is an L_2 function by showing that the individual terms are L_2. Define:

$$M_f(v) = \int_{-\infty}^{+\infty} e^{-jvx} f(x)dx$$

$$M_\alpha(v) = \int_{-\infty}^{+\infty} e^{-jvz} d\alpha(z) .$$

Recall that X = N + Z, and that N and Z are independent. Then,

$$M_f(v) = e^{-\frac{1}{2}\ \sigma^2v^2}\ M_\alpha(v) . \qquad (2.4.20)$$

Clearly, $v^2M_f(v)$ is L_2. Then from Plancherel's theorem its transform $f''(x)$ is L_2. Similarly, if $\frac{d^2}{dv^2} M_f(v)$ is L_2, $x^2f(x)$ will be L_2. With the second moment of Z finite, the first and second derivatives of $M_\alpha(v)$ exist, are continuous and bounded. From this it follows that

$\frac{d^2}{dv^2} M_f(v) \in L_2$. In this manner, all terms of $w(x)$ are seen to be L_2 functions. q.e.d.

It should be clear as to the conditions needed to guarantee faster convergence of the MISE. Assume that the r-th absolute moment of z exists, $\int_{-\infty}^{+\infty} |z|^r d\alpha(z) < \infty$. By repeated application of the method of the lemma we obtain:

$$a_j^2 \leq B_6/j^r$$

$$\sum_{j=q+1}^{\infty} a_j^2 \leq B_6 \left(\frac{1}{(r-1)q^{r-1}} + \frac{1}{q^r} \right). \qquad (2.4.21)$$

B_6 is now the L_2 norm of the function

$$\frac{\sigma_1^{2r}}{g(x/\sigma_1)} \frac{1}{2^{r/2}} \frac{d^r}{dx^r} (g(x/\sigma_1)f(x)).$$

We summarize this discussion in

Theorem 2.4.2: Under the hypotheses of Lemma 2.4.1, the sequence of estimates

$$\hat{f}_n(X_1, X_2, \ldots X_n; x) = \sum_{j=0}^{q(n)} \hat{a}_{jn} \varphi_j(x/\sigma_1) \qquad (2.4.11)$$

with $q(n) \to \infty$ and $q(n)/n \to 0$ converge in the sense of MISE to $f(x)$. If

$$\int |z|^r d\alpha(z) < \infty, \quad r \geq 2,$$

the MISE at the n-th stage satisfies the inequality

$$J_n = E\int (f_n(x) - f(x))^2 dx \leq B_6\left(\frac{1}{(r-1)q(n)^{r-1}} + \frac{1}{q(n)^r}\right) + \frac{c_3 q(n)}{n} \tag{2.4.22}$$

where c_3 is defined in (10). Choosing $q(n)$ as the largest integer $\leq (B_6 n/c_3)^{1/r}$ gives

$$J_n = O(1/n^{(r-1)/r}) . \tag{2.4.23}$$

We now want to consider the mean-square error for fixed x. With the estimate $\hat{f}_n(x)$ as in (14), define the function

$$f_q(x) = \sum_{j=0}^{q(n)} a_j \varphi_j(x/\sigma_1) . \tag{2.4.24}$$

The mean-square error, as a function of x, is given by

$$E\left\{(f(x) - \hat{f}_n(X_1,X_2,\ldots X_n;x))^2\right\} = E\left\{(f(x) - \hat{f}_n(x))^2\right\} =$$

$$(f(x) - f_q(x))^2$$

$$+ 2(f(x) - f_q(x)) \sum_{j=0}^{q} E(a_j - \hat{a}_{jn})\, \varphi_j(x/\sigma_1)$$

$$+ \sum_{\substack{j=0 \\ k=0}}^{q} E\left((a_j - \hat{a}_{jn})(a_k - \hat{a}_{kn})\right) \varphi_j(x/\sigma_1)\varphi_k(x/\sigma_1) . \tag{2.4.25}$$

Since $\hat{a}_{jn}$ is an unbiased estimate, the cross-product term is zero. The other expectation of (25) is bounded by Schwarz's inequality and Theorem 2.4.1:

$$E\left\{(a_j - \hat{a}_{jn})\,(a_k - \hat{a}_{kn})\right\} < \left[V(\hat{a}_{jn})\, V(\hat{a}_{kn})\right]^{1/2} < \frac{c_3}{n} .$$

Using (7), the third expression in (25) is dominated by $c_2^2 q^2(n) c_3/n$. Hence, we obtain the bound:

$$E\left\{\left[(f(x)-f_n(x))^2\right]\right\} \leq (f(x)-f_q(x))^2 + c_2^2\, c_3\, \frac{q^2(n)}{n} . \qquad (2.4.26)$$

To continue the discussion we need to investigate the pointwise convergence of $f_q(x)$ to $f(x)$.

Theorem 2.4.3 (Sansone [38], p. 381): If $f(x)$ is contained in L_1 and L_2 then, at a finite point x_0, the series

$$\sum_{j=0}^{\infty} a_j\, \varphi_j(x/\sigma_1)$$

behaves like the Fourier trigonometric series of a function which coincides with $f(x)$ in an arbitrarily small neighborhood (x_0-h, x_0+h) of x_0.

In particular if $f(x)$ is of bounded variation in a neighborhood of x_0 we have

$$\sum_{j=0}^{\infty} a_j\, \varphi_j(x_0/\sigma_1) = \frac{1}{2}\left[f(x_0^+) + f(x_0^-)\right] ,$$

and again if $f(x)$ is continuous and of bounded variation in $(-\infty,\infty)$, then the series converges uniformly in any interval interior to $(-\infty,\infty)$.

Sansone's proof, which is attributed to J. V. Uspensky, involves showing that in a neighborhood of x_0 the partial sum of the series can be made arbitrarily close to the partial sum of the Fourier expansion of the function in the same interval. This adaptation of Fejer's method

of Fourier series is also used by Wiener to obtain the same result.[1] The above theorem is a special case of the more general result that for orthogonal functions of the Sturm-Liouville type, the L_2 series of these functions behaves in the same manner at a point as Fourier's series do (Hobson [18], p. 771).

The density function $f(x)$ is L_1, L_2, and continuous for every x. It is also of bounded variation in any interval since $f'(x)$ exists and is bounded (see (2.3.12)). Hence, in any finite interval, as $n \to \infty$ we obtain

$$f_q(x) = \sum_{j=0}^{q(n)} a_j \varphi_j(x/\sigma_1) \longrightarrow f(x), \text{ uniformly in } x.$$

Specifying a rate again involves a moment assumption. Assuming the r-th absolute moment of Z exists, where now $r \geq 3$ (cf. Theorem 2.4.2), yields:

$$\begin{aligned} |f(x)-f_q(x)| &\leq \Big| \sum_{j=q+1}^{\infty} a_j \varphi_j (x/\sigma_1) \Big| \\ &\leq c_2 \sum_{j=q+1}^{\infty} |a_j| \\ &\leq c_2 \sqrt{B_6} \sum_{j=q+1}^{\infty} \frac{1}{j^{r/2}} \\ &\leq c_2 \sqrt{B_6} \left(\frac{1}{(\frac{r}{2}-1)\, q^{\frac{r}{2}-1}} + \frac{1}{q^{\frac{r}{2}}} \right). \end{aligned} \tag{2.4.27}$$

[1] N. Wiener, The Fourier Integral and Certain of Its Applications, Dover Publications, Inc., N.Y., 1933, pp. 55-67.

We have used (7), the first part of (21), and the usual integral upper bound. We are now in a position to prove

Theorem 2.4.4: Assuming the hypotheses of Lemma 2.4.1 hold, then, since f(x) satisfies the conditions of the previous theorem, the sequence of estimates

$$\hat{f}_n(x) = \sum_{j=1}^{q(n)} \hat{a}_{jn}\varphi_j(x/\sigma_1) \tag{2.4.11}$$

converge in mean-square to f(x) if $q \to \infty$ and $q^2/n \to 0$. This convergence is uniform in x for any finite interval. If the r-th absolute moment of z exists, $r \geq 3$, the mean-square error is dominated by

$$E\left\{(f(x)-\hat{f}_n(x))^2\right\} \leq B_6 c_2^2 \left(\frac{1}{(\frac{r}{2}-1)q^{\frac{r}{2}-1}} + \frac{1}{q^{\frac{r}{2}}}\right)^2 + \frac{c_3 c_2^2 q^2(n)}{n} . \tag{2.4.28}$$

Upon choosing $q(n) = [n^{1/r}]$, as $n \to \infty$ we obtain

$$E\left\{(f(x) - \hat{f}_n(x))^2\right\} = O(1/n^{\frac{r-2}{r}}) .$$

For the application of the empirical Bayes technique, we shall need the convergence of $\hat{f}_n(X_n)$ to the random variable $f(X_n)$. The mean-square error in this case is written (see (25)):[1]

$$E_n\left\{(f(X_n) - \hat{f}_n(X_1,\ldots X_n;X_n))^2\right\} =$$

$$E_n\left\{(f(X_n) - \hat{f}_n(X_n))^2\right\} =$$

[1] We use the E_n notation of Chapter 1.

$$= E_n\left\{(f(X_n) - f_q(X_n))^2\right\}$$

$$+ 2\,E_n\left\{(f(X_n) - f_q(X_n))\sum_{j=0}^{q(n)} (a_j - \hat{a}_{jn})\,\varphi_j(X_n/\sigma_1)\right\}$$

$$+ E_n\left\{\sum_{\substack{j=0\\k=0}}^{q(n)} (a_j - \hat{a}_{jn})(a_k - \hat{a}_{kn})\varphi_j(X_n/\sigma_1)\varphi_k(X_n/\sigma_1)\right\}. \qquad (2.4.29)$$

Now, the first and third terms have already been bounded independent of the argument. With a r-th absolute moment assumption on the random variable Z, from (27) we have

$$E_n\left\{(f(X_n) - f_q(X_n))^2\right\} \le \left(c_2\sqrt{B_\theta}\left(\frac{1}{(\frac{r}{2}-1)q^{\frac{r}{2}-1}} + \frac{1}{q^{\frac{r}{2}}}\right)\right)^2,$$

and as before, the third term of (29) is dominated by $\frac{c_3}{n} c_2^2 q^2$. Using these bounds and the Minkowski inequality, (29) is dominated by

$$E_n\left\{(f(X_n) - \hat{f}_n(X_n))^2\right\} \le \left\{c_2\sqrt{B_\theta}\left(\frac{1}{(\frac{r}{2}-1)q^{r/2-1}} + \frac{1}{q^{r/2}}\right) + c_2\, q\,\sqrt{\frac{c_3}{n}}\right\}^2. \qquad (2.4.30)$$

The fastest rate of convergence is obtained if q(n) is set equal to the largest integer less than or equal to $(B_\theta n/c_3)^{1/r}$. This gives:

$$E_n\left\{(f(X_n) - \hat{f}_n(X_n))^2\right\} \le c_2\sqrt{B_\theta}\left(\frac{1}{(\frac{r}{2}-1)\left[(\frac{B_\theta}{c_3}n)^{1/r} - 1\right]^{\frac{r}{2}-1}}\right.$$

$$\left. + \frac{1}{\left[(\frac{B_\theta}{c_3})^{1/r} - 1\right]^{\frac{r}{2}}} + \frac{c_2\sqrt{c_3}\,(B_\theta/c_3)^{1/r}}{n^{(r-2)/2r}}\right)^2 \qquad (2.4.31)$$

Asymptotically, (31) is dominated by

$$E_n\left\{(f(X_n)-\hat{f}_n(X_n))^2\right\} \leq \frac{1}{n^{(r-2)/r}}\left\{\frac{c_2}{(\frac{r}{2}-1)}\ c_3^{\frac{1}{r}}\ B_6^{\frac{r-2}{2r}} + c_2\ B_6^{\frac{1}{r}}\ c_3^{\frac{r-2}{2r}}\right\}^2 . \qquad (2.4.32)$$

Corollary 2.4.1: Under the hypotheses of Lemma 2.4.1 and with the r-th absolute moment of Z finite, $r \geq 3$, $\hat{f}_n(X_n)$ converges in mean-square to $f(X_n)$. The mean-square error is dominated by (31), and for large n we have

$$E_n\left\{(f_n(X_n)-\hat{f}_n(X_n))^2\right\} = O(1/n^{(r-2)/r}). \qquad (2.4.33)$$

It is not surprising of course, that we achieve the same rate as in Theorem 2.4.4.

In practice, we may want to hold q fixed, i.e., estimate an approximation of f(x). We then take the estimate

$$\tilde{f}_n(x) = \sum_{j=0}^{q} \hat{a}_{jn}\ \varphi_j(x/\sigma_1) . \qquad (2.4.34)$$

q is now a fixed integer chosen according to some error criterion.

The estimates of the q + 1 coefficients, (4), can be put in the recursive form

$$\hat{a}_{jn} = \frac{1}{n}\left[(n-1)\ \hat{a}_{jn-1} + \varphi_j(X_n/\sigma_1)\right],\ j = 0,1,\ldots q. \qquad (2.4.35)$$

$\tilde{f}_n(x)$ converges in MISE to a function which differs from f(x) in

the L_2 sense by $\sum_{q+1}^{\infty} a_j^2$, i.e., (12) gives

$$E \int (f(x)-\tilde{f}_n(x))^2 dx \leq \sum_{j=q+1}^{\infty} a_j^2 + \frac{c_3 q}{n} \quad . \tag{2.4.36}$$

In practice, it can reasonably be assumed that the second moment of Z exists. Hence, the asymptotic error in the estimate is then known to be inversely proportional to the number of terms used in the series ((21)).

Similarly, we can take $\tilde{f}_n(X_n)$ as an approximate estimate of $f(X_n)$. With r greater than two, the mean-square error ((30)) is bounded by

$$E_n\left\{(f(X_n)-\tilde{f}_n(X_n))^2\right\} \leq \left\{c_2\sqrt{B_6}\left(\frac{1}{(\frac{r}{2}-1)\, q^{\frac{r}{2}-1}} + \frac{1}{q^{\frac{r}{2}}}\right) + \frac{c_2 q \sqrt{c_3}}{\sqrt{n}}\right\}^2 . \tag{2.4.37}$$

The asymptotic error is

$$\lim_{n\to\infty} E_n\left\{(f(X_n)-\tilde{f}_n(X_n))^2\right\} = O(1/q^{r-2}) ,$$

which is inversely proportional to some power of the number of terms used.

2.5 ESTIMATING THE DENSITY BY SERIES METHODS—AN EIGENFUNCTION REPRESENTATION

The method we now discuss makes use of the fact that f(x) is the convolution of a known gaussian density function with an unknown distribution. This is in contradistinction to the previous methods which do not utilize this knowledge.

In the equation for f(x),

$$f(x) = \int_{-\infty}^{+\infty} g(x-z;\sigma)\, d\alpha(z) ,$$

$\alpha(z)$ is the unknown quantity. To make use of this we want to express $f(x)$ is a form which, in some sense, isolates $\alpha(z)$. This is accomplished by solving an eigenfunction problem associated with the above equation.

We first observe that the "kernel" $g(x-z;\sigma)$ is not Hilbert-Schmidt; it is not square integrable in the x-z product space. We will display a function $s(x)$ such that

$$\int_x \int_z s^2(x)\, g^2(x-z;\sigma)dx\, dz < \infty \ .$$

A considerably more difficult task is to choose a $s(x)$ so that we can solve for the eigenfunctions and eigenvalues of the operator $s(x)g(x-z;\sigma)$, i.e., find the φ's and λ's which satisfy

$$\int_{-\infty}^{+\infty} s(x)g(x-z;\sigma)\ \varphi_j(z/\sigma_1)dz \ = \ \lambda_j\ \varphi_j(x/\gamma) \ .$$

We shall find these quantities by obtaining the diagonal L_2 expansion

$$s(x)g(x-z;\sigma) = \sum_{j=0}^{\infty} \lambda_j\ \varphi_j(x/\gamma)\ \varphi_j(z/\sigma_1) \ .$$

Note that by a change of variables $y = \frac{x\sigma_1}{\gamma}$, the right side of the expansion is symmetrical in y and z. Hence, the operator $s(\gamma y/\sigma_1)g(\frac{\gamma y}{\sigma_1} - z;\sigma)$ is also symmetrical and the above formulae reduce to more familiar forms. For our purposes, it is more convenient to deal with the unsymmetrical operator, $s(x)g(x-z;\sigma)$.

Having found the φ's, we define the coefficients associated with the distribution $\alpha(z)$:

$$d_j = \int_{-\infty}^{+\infty} \varphi_j(z/\sigma_1)\, d\alpha(z)$$

Under suitable conditions, the quantity $s(x)f(x)$ can be written:

$$s(x)f(x) = \int_{-\infty}^{+\infty} s(x)\, g(x-z;\sigma) d\alpha(z)$$

$$= \sum_{j=0}^{\infty} \lambda_j d_j \varphi_j(x/\gamma)$$

It is this form which displays the unknown quantities—the d_j's.

Consequently, we will estimate not $f(x)$, but rather, the product $s(x)f(x)$. Since $s(x)$ turns out to be a positive function, it is then just a matter of dividing by $s(x)$ to discuss the mean-square error. However, as discussed in section 1.2, when an equivalent test which incorporates $s(x)$ can be found, the quantity we are interested in estimating is just the product $s(x)f(x)$. We proceed to obtain the above expansion.

Consider the gaussian density appearing in the convolution. $g(x-z;\sigma)$ is an L_2 function in z for every x. Expand this function in the orthonormal series

$$g(x-z;\sigma) = \sum_{j=0}^{\infty} c_j(x,\sigma,\sigma_1)\varphi_j(z/\sigma_1) \qquad (2.5.1)$$

where the φ_j are again the Hermite functions as in (2.4.3), and σ_1 is an arbitrary positive constant. Here the expansion is in terms of the independent variable z. As indicated, the coefficients are functions

of x, σ_1 and σ, and are calculated by

$$c_j(x,\sigma_1,\sigma) = \int_{-\infty}^{+\infty} g(x-z;\sigma)\ \varphi_j(z/\sigma_1)dz\ . \tag{2.5.2}$$

The evaluation of this integral is the main result of Appendix B. Make the definitions:

$$\begin{aligned} \xi^2 &= \left\{\frac{\sigma_1^2-\sigma^2}{\sigma_1^2+\sigma^2}\right\} \\ \gamma^2 &= \frac{(\sigma_1^2-\sigma^2)(\sigma_1^2+\sigma^2)}{\sigma_1^2} \end{aligned} \tag{2.5.3}$$

Then, from (B.10),

$$c_j(x_1\sigma_1,\sigma) = \xi^j\ \frac{g(x;\sqrt{\sigma_1^2+\sigma^2})\ H_j(x/\gamma)}{\sqrt{2^j j!/\sqrt{4\pi}\,\sigma_1}} \tag{2.5.4}$$

and hence,

$$g(x-z;\sigma) = g(x;\sqrt{\sigma_1^2+\sigma^2})\sum_{j=0}^{\infty}\ \frac{\xi^j\ H_j(x/\gamma)\varphi_j(z/\sigma_1)}{\sqrt{2^j j!/\sqrt{4\pi}\,\sigma_1}} \tag{2.5.5}$$

Now consider the function

$$s(x) = \sqrt{\frac{\gamma}{\sigma_1}}\ \frac{g(x;\gamma)}{g(x;\sqrt{\sigma_1^2+\sigma^2})}$$

$$= \left\{\frac{\sigma_1^2+\sigma^2}{\sigma_1^2-\sigma^2}\right\}^{1/4} \exp\left(-\frac{x^2}{2}\left\{\frac{\sigma^2}{(\sigma_1^2+\sigma^2)(\sigma_1^2-\sigma^2)}\right\}\right) = \frac{1}{\sqrt{\xi}}\ \exp\left(-\frac{x^2}{2}\,\frac{\sigma^2}{\sigma_1^2\gamma^2}\right) \tag{2.5.6}$$

Let σ_1 be greater than σ but otherwise arbitrary. Then, s(x) is a bounded function. It is also non-negative, L_1, and L_2. Multiply (5) by s(x)

$$s(x)g(x-z;\sigma) = \sum_{j=0}^{\infty} \xi^j \varphi_j(x/\gamma)\varphi_j(z/\sigma_1) \quad . \tag{2.5.7}$$

$s(x)$ is just the right function to make the set of coefficients $c_j(x,\sigma)$ an orthogonal set (with respect to dx); or what is the same thing, (7) is an L_2 expansion in the x-z product space with the coefficient $a_{ij} = 0$ if $i \neq j$, $a_{jj} = \xi^j$.[1] In fact, we even have more. Since $g(x-z;\sigma)$ as a function of z satisfies the conditions of the theorem quoted earlier (Theorem 2.4.3), (5) converges uniformly in z for every x. Hence, we have pointwise convergence in x and z. Multiplying by $s(x)$ does not change this convergence and (7) converges pointwise to $s(x)g(x-z;\sigma)$. Since $\xi < 1$, it is easy to see that the error in the remainder term of (7) is dominated by

$$\left| \sum_{j=J}^{\infty} \xi^j \varphi_j(x/\gamma)\varphi_j(z/\sigma_1) \right| \leq \frac{c_1^2}{\sqrt{\pi\gamma\sigma_1}} \frac{\xi^{J+1}}{1-\xi} \quad . \tag{2.5.8}$$

Since the remainder is independent of x and z, (7) converges uniformly in x and z.

Now write

$$s(x)f(x) = \int_{-\infty}^{+\infty} s(x)g(x-z;\sigma)d\alpha(z)$$

and substitute (7).

$$s(x)f(x) = \int_{-\infty}^{+\infty} \left[\sum_{j=0}^{\infty} \xi^j \varphi_j(x/\gamma)\, \varphi_j(z/\sigma_1) \right] d\alpha(z) \quad . \tag{2.5.9}$$

[1] It is not difficult to calculate the L_2 norms for both sides of (7). It is equal to $\sum (\xi^j)^2 = (\sigma_1^2+\sigma^2)/2\sigma^2$.

Define d_j as the j-th coefficient associated with the distribution $\alpha(z)$.

$$d_j = \int_{-\infty}^{+\infty} \varphi_j(z/\sigma_1)d\alpha(z) \tag{2.5.10}$$

We justify interchanging the operations in (9) by the Lebesgue dominated convergence theorem. The result is:

$$s(x)f(x) = \sum_{j=0}^{\infty} \xi^j \, d_j \, \varphi_j(x/\gamma) \tag{2.5.11}$$

Equation (11) is an orthogonal representation for $s(x)f(x)$. It also follows, in a number of ways, that the series converges pointwise. In view of the inequalities

$$|d_j| \leq \int |\varphi_j(z/\sigma_1)|d\alpha(x) \leq$$

$$\frac{c_1}{(\pi^{1/4} \sigma_1^{1/2})} \int d\alpha(z) = \frac{c_1}{(\pi^{1/4} \sigma_1^{1/2})} \tag{2.5.12}$$

and $|\xi| < 1$, it also follows that

$$|s(x)f(x) - \sum_{j=0}^{q} \xi^j \, d_j \, \varphi_j(x/\gamma)|$$

$$\leq \frac{c_1^2}{\sqrt{\pi\sigma_1\gamma}} \frac{\xi^{q+1}}{1-\xi} \tag{2.5.13}$$

and the series in (11) converges uniformly in x.

To estimate the quantity $s(x)f(x)$, we proceed in a manner analogous to section 2.4. Take, as estimates of the product of the coefficients

$\xi^j d_j$, the quantity

$$\xi^j \hat{d}_{jn} = \frac{1}{n} \sum_{\ell=1}^{n} \varphi_j(X_\ell/\gamma)\, s(X_\ell),\ j = 1,2,\ldots . \tag{2.5.14}$$

The estimates are unbiased:

$$\begin{aligned} E(\xi^j \hat{d}_{jn}) &= E(\varphi_j(X/\gamma)s(X)) \\ &= \int \varphi_j(x/\gamma)s(x)f(x)dx \\ &= \xi^j d_j . \end{aligned} \tag{2.5.15}$$

Since the function $\varphi_j(x/\gamma)s(x)$ is bounded (uniformly in j and x) by

$$\varphi_j(x/\gamma)s(x) \le \frac{c_1}{\pi^{1/4}\gamma^{1/2}} \left\{\frac{\sigma_1^2+\sigma^2}{\sigma_1^2-\sigma^2}\right\}^{1/4} = \frac{c_1}{\pi^{1/4}} \frac{1}{(\gamma\xi)^{1/2}}$$

$$= \frac{c_1}{\pi^{1/4}} \frac{\sigma_1}{(\sigma_1^2-\sigma^2)^{3/4}(\sigma_1^2+\sigma^2)^{1/4}} = c_4 , \tag{2.5.16}$$

in analogy to Theorem 2.4.1 (see (2.4.10)) we have

Theorem 2.5.1: Under the hypotheses of Lemma 2.4.1, the estimates $\xi^j \hat{d}_{jn}$ converge in mean-square at the rate 1/n:

$$E\left\{\xi^{2j}(d_j-\hat{d}_{jn})^2\right\} = V(\xi^j\hat{d}_{jn}) \le \frac{c_4^2}{n}\left(2 + \frac{2B}{1-\rho^*} + 4(M-1)\right) = \frac{c_5}{n} \tag{2.5.17}$$

As the estimate of $s(x)f(x)$ at the n-th observation we take

$$s(x)\hat{f}_n(x) = \sum_{j=0}^{q(n)} \xi^j \hat{d}_{jn}\, \varphi_j(x/\gamma) . \tag{2.5.18}$$

Designating J_n' as the MISE, we have:

$$J_n' = E \int (s(x)(f(x)-\hat{f}_n(x)))^2 \, dx$$

$$= \sum_{j=q+1}^{\infty} \xi^{2j} d_j^2 + \sum_{j=1}^{q} E\left\{\xi^{2j} (\hat{d}_{jn}-d_j)^2\right\} \qquad (2.5.19)$$

In view of (12) and the previous theorem, J_n' is dominated by

$$J_n' = \frac{c_1^2}{\sqrt{\pi}\,\sigma_1} \frac{\xi^2}{1-\xi^2} \xi^{2q(n)} + \frac{c_5 q(n)}{n} . \qquad (2.5.20)$$

For fixed n, this bound is minimized if q(n) is chosen as

$$q(n) = \frac{1}{2\,\ell n \xi} \ell n \left\{\frac{c_5 \sigma_1 \sqrt{\pi}}{c_1^2 \ell n\, \xi^2}\right\} - \frac{\ell n\, n}{2\ell n \xi} = c_6 - \frac{\ell n\, n}{2\ell n \xi} . \qquad (2.5.21)$$

The logarithm is taken to the base e. Since $0 < \xi < 1$, $q(n) \to \infty$. Letting q(n) equal the largest integer less than or equal to (21), J_n' is dominated by

$$J_n' \leq \left(\frac{c_1^2}{\pi^{1/2}\sigma_1}\right) \frac{\xi^{2c_6}}{1-\xi^2} \frac{1}{n} + \frac{c_5}{n}\left(c_6 + \frac{\ell n\, n}{2|\ell n \xi|}\right). \qquad (2.5.22)$$

Theorem 2.5.2: Under the hypotheses of Lemma 2.4.1, the sequence of estimates

$$s(x)\hat{f}_n(x) = \frac{1}{n} \sum_{\ell=1}^{q(n)} \xi^j \hat{d}_{jn} \varphi_j(x/\gamma)$$

with coefficients

$$\xi^j \hat{d}_{jn} = \frac{1}{n} \sum_{\ell=1}^{n} \varphi_j(X_\ell/\gamma) s(X_\ell)$$

converge in MISE to $s(x)f(x)$ if $q(n)$ is chosen as above. In this case, the MISE satisfies

$$J_n' = O\left(\frac{\ell n\ n}{n}\right) \tag{2.5.23}$$

Observe that no assumptions on the moments of the random variable Z are required.

The mean-square error for fixed x is given by (cf.(2.4.26)):

$$\begin{aligned} E(s^2(x)(f(x)-\hat{f}_n(x))^2) &= s^2(x)(f(x)-f_q(x))^2 \\ &+ 2s(x)(f(x)-f_q(x))\sum_{j=0}^{q} E(\xi^j(d_j-\hat{d}_{jn}))\varphi_j(x/\gamma) \\ &+ \sum_{\substack{j=0\\k=0}}^{q} E(\xi^j\xi^k(d_j-\hat{d}_{jn})(d_k-\hat{d}_{kn}))\varphi_j(x/\gamma)\varphi_k(x/\gamma) \ . \end{aligned} \tag{2.5.24}$$

$f_q(x)$ is defined implicitly by the equation

$$s(x)f_q(x) = \sum_{j=0}^{q(n)} \xi^j d_j\ \varphi_j(x/\gamma) \ . \tag{2.5.25}$$

From Theorem 2.5.1 and (2.4.7), the third term of (24) is dominated by

$$\frac{c_1^2}{\sqrt{\pi}\ \sigma} \qquad \frac{c_5\ q^2(n)}{n}$$

The first term of (24) is majorized using (13), and the middle term is zero. Combining bounds yields:

$$E(s^2(x)\ (f(x)-\hat{f}_n(x))^2) \leq \frac{c_1^4}{(\pi\sigma_1\gamma)}\ \frac{\xi^2}{(1-\xi)^2}\ \xi^{2q(n)} + \frac{c_1^2}{\sqrt{\pi}\,\gamma}\ \frac{c_5\ q^2(n)}{n}\ . \tag{2.5.26}$$

The fastest rate of convergence is obtained if q(n) is set equal to the largest integer less than or equal to

$$\frac{\ell n\ n}{2\ \ell n\ \xi}\ .$$

Using the inequality $\frac{\ell n(n)}{2\ \ell n\ \xi} - 1 \leq q(n) \leq \frac{\ell n(n)}{2\ \ell n\ \xi}$, we have:

$$E\left\{s^2(x)\ (f(x)-\hat{f}_n(x))^2\right\} \leq \frac{c_1^4}{(\pi\sigma_1\gamma)(1-\xi)^2}\ \frac{1}{n} + \frac{c_1^2 c_5}{4\sqrt{\pi}\ \gamma}\ \frac{1}{n}\ \left(\frac{\ell n\ n}{\ell n\ \xi}\right)^2\ . \tag{2.5.27}$$

Theorem 2.5.3: Under the hypotheses of Lemma 2.4.1, the mean-square error (for fixed x) satisfies

$$E\left\{s^2(x)(f(x)-\hat{f}_n(x))^2\right\} = O\left(\frac{\ell n^2 n}{n}\right) \tag{2.5.28}$$

Consequently, the sequence of estimates

$$\hat{f}_n(x) = \frac{1}{s(x)}\ \frac{1}{n}\ \sum_{j=0}^{q(n)}\ \xi^j\ \hat{d}_{jn}\ \varphi_j(x/\gamma) \tag{2.5.29}$$

converge in mean-square to f(x) at the same rate. This convergence is uniform in any finite interval.

To investigate the mean-square convergence of $s(X_n)\hat{f}_n(X_n)$ to the random variable $s(X_n)f(X_n)$ we write:

$$E_n\left\{(s^2(X_n)(f(X_n)-\hat{f}_n(X_n))^2\right\} =$$

$$E_n\left\{s^2(X_n)(f(X_n)-f_q(X_n))^2\right\}$$

$$+ 2E_n\left\{s(X_n)(f(X_n)-f_q(X_n))\sum_{j=0}^{q(n)} \xi^j (d_j-\hat{d}_{jn})\varphi_j(X_n/\gamma)\right\}$$

$$+ E_n\left\{\sum_{\substack{j=0\\k=0}}^{q(n)} \xi^j\xi^k (d_j-\hat{d}_{jn})(d_j-\hat{d}_{jn})\varphi_j(X_n/\gamma)\varphi_k(X_n/\gamma)\right\} \qquad (2.5.30)$$

The first and third terms of this equation have just been bounded independent of the argument. From the Minkowski inequality, (13), and the expression below (25), we have (cf.(26)):

$$E_n\left\{s^2(X_n)(f(X_n)-\hat{f}_n(X_n))^2\right\} \leq$$

$$\left\{\frac{1}{\sqrt{\pi\sigma_1\gamma}}\,\frac{c_1^2\xi}{(1-\xi)}\,\xi^{q(n)} + c_1\sqrt{\frac{c_5}{\sqrt{\pi}\gamma}}\,\frac{q(n)}{\sqrt{n}}\right\}^2 . \qquad (2.5.31)$$

This expression is minimized for fixed n if q(n) is chosen as

$$q(n) = c_7 + \frac{\ell n(n)}{2|\ell n\xi|} \qquad (2.5.32)$$

where

$$c_7 = \frac{\pi^{1/4}(\sigma_1 c_5)^{1/2}(1-\xi)}{c_1\xi} \qquad (2.5.33)$$

Setting q equal to the largest integer less than or equal to (32), we have:[1]

$$E_n\left\{ s^2(X_n)(f(X_n)-\hat{f}_n(X_n))^2 \right\} \leq$$
$$\left\{ \frac{c_1{}^2 \xi^{c_7}}{\sqrt{\pi\sigma_1\gamma}(1-\xi)} \frac{1}{\sqrt{n}} + c_1 \sqrt{\frac{c_5}{\sqrt{\pi}\gamma}} \left(c_7 + \frac{\ell n\,(n)}{2\,\ell n|\xi|}\right) \frac{1}{\sqrt{n}} \right\}^2 . \tag{2.5.34}$$

Corollary 2.5.1: Under the hypotheses of Lemma 2.4.1, the sequence of estimates

$$s(X_n)\hat{f}_n(X_n) = \frac{1}{n} \sum_{j=0}^{q(n)} \xi^j \hat{d}_{jn} \varphi_j(X_n/\gamma) ,$$

with q(n) given by (32), converge in mean-square to the random variable $s(X_n)f(X_n)$. The mean-square error is dominated by (34). Hence, we achieve the rate:

$$E_n\left\{ s^2(X_n)(f(X_n)-\hat{f}_n(X_n))^2 \right\} = O\left(\frac{\ell n^2(n)}{n}\right) . \tag{2.5.35}$$

There are a number of differences between this method of estimation and the previous L_2 series representation. To apply the method of this section the standard deviation of the noise must be known while in the previous method a moment assumption on the random variable Z is needed to specify a rate. Using the present method, we have obtained a rate of

[1] We have used $q \geq c_7 - 1 + \frac{\ell n(n)}{2|\ell n\xi|}$ in the first expression on the right side of (34).

$\ln^2 n/n$ for the mean-square error (Thm.2.5.3) and $\ln(n)/n$ for the MISE (Thm. 2.5.2). For the first series method, with the r-th absolute moment of Z finite, we had the rate $1/n^{(r-2)/r}$ for the mean-square error (Thm. 2.4.4) and $1/n^{(r-1)/r}$ for the MISE.[1]

In practice, we may want to hold q fixed, settling for an approximate estimate of $s(x)f(x)$. In this case, we designate the estimate by

$$s(x)\tilde{f}_n(x) = \sum_{j=0}^{q} \xi^j \hat{d}_{jn} \varphi_j(x/\gamma)$$

Considering the MISE we have (see (20)):

$$\lim_{n\to\infty} E \int (s^2(x)(f(x)-\tilde{f}_n(x))^2 \, dx =$$

$$= \sum_{j=q+1}^{\infty} \xi^{2j} d_j^2 \leq \frac{c_1^2}{\sqrt{\pi}\,\sigma_1} \frac{\xi^2}{1-\xi^2} \xi^{2q} = \frac{c_1^2}{\sqrt{\pi}} \frac{\sigma_1}{2\sigma^2} \left\{\frac{\sigma_1^2-\sigma^2}{\sigma_1^2+\sigma^2}\right\}^{q+\frac{1}{2}} \tag{2.5.36}$$

Similarly, $s(X_n)\tilde{f}_n(X_n)$ converges in mean-square to $s(X_n)f_q(X_n)$. The mean-square error is given by (31) and the asymptotic error is:

$$\lim_{n\to\infty} E_n\left\{s^2(X_n)(f(X_n)-\tilde{f}_n(X_n))^2\right\}$$

$$= \frac{c_1^4 \xi^2}{(\pi \sigma_1 \gamma)(1-\xi)^2} \xi^{2q}$$

---

[1] The comparison of rates for MISE is not really valid as different quantities are being estimated.

$$\leq \frac{c_1^4}{\pi} \frac{1}{4\sigma^4} \frac{(\sigma_1^2-\sigma^2)^{2q+1}}{(\sigma_1^2+\sigma^2)^{2q+3}} \quad . \tag{2.5.37}$$

The asymptotic error in both cases decreases geometrically with q. In the previous method, the asymptotic error is inversely proportional to some power of q.

Another significant difference between the two methods is when we want to estimate the density function $\alpha'(z)$. This will be discussed in section 4.4.

Recall that σ_1 is an arbitrary constant chosen to be greater than σ. (σ is the standard deviation of the noise samples.) For applications in hypothesis testing with a minimum probability of error criterion, we will use a test function of the form

$$p_o s(x) f_o(x) - p_1 s(x) f_1(x)$$

Since $s(x)$ is proportional to

$$\exp\left\{-\frac{x^2}{2}\left\{\frac{\sigma^2}{(\sigma_1^2-\sigma^2)(\sigma_1^2+\sigma^2)}\right\}\right\} \quad ,$$

the free parameter σ_1 can be considered as a scaling factor for the test functional. We shall discuss the problem of choosing σ_1 in section 4.1.

We remark that in the series method of section 2.4, σ_1 is also arbitrary and would naturally be chosen to minimize the bound on convergence. For example, it enters into the MISE bound through two terms: B_6 is proportional to σ_1^{2r} and c_3 is proportional to $1/\sigma_1$.

2.6 SPECIAL FORMS OF $\alpha(z)$

We consider the case where $\alpha(z)$ contains a finite set of unknown parameters which enter linearly into $f(x)$. Examples are equations (1.3.5) and (1.3.7) with the set of a priori probabilities unknown. We take (1.3.5)

$$\alpha(z) = \sum_{i=1}^{k} p_i \, u(z-y_i) \tag{1.3.5}$$

where $u(z)$ is the unit step function. Then,

$$f(x) = \int_{-\infty}^{+\infty} g(x-z;\sigma)\,d\alpha(z) = \sum_{i=1}^{k} p_i \, g(x-y_i;\sigma) \, , \tag{1.3.6}$$

and the problem is to find a sequence of estimates, $\hat{p}_{i,n}$, which converge to the p_i for $i=1,2,\ldots K$. For the case of independent samples, this problem has been solved.[1] The procedure used is still applicable in our situation.

A necessary and sufficient condition for the existence of the sequence $\hat{p}_{i,n}$ is that the signals (the mean values of the gaussian density functions) y_i, $i=1,2,\ldots K$, be distinct. The condition is clearly necessary for if $y_i = y_j$ we can not distinguish between the hypotheses i,j, or the a priori probabilities p_i, p_j. Sufficiency is demonstrated by constructing the sequence.

---

[1] Robbins [31]. For a discussion of this general type of problem see H. Teicher, "Identifiability of Finite Mixtures," Ann. Math. Stat., v. 34, 1963, pp. 1265-1269.

We will show the condition that the signals be distinct is equivalent to having the K functions $g(x-y_i;\sigma)$, i=1,...K, linearly independent. Postponing this proof till later, we now give a procedure for estimating the p_i. Assuming that the $g(x-y_i;\sigma)$ are linearly independent, define:

$$g_{ij} = \int_{-\infty}^{+\infty} g(x-y_i;\sigma)g(x-y_j;\sigma)dx = g(y_i-y_j;\sqrt{2}\,\sigma) \qquad (2.6.1)$$

G = the Gramian matrix whose elements are g_{ij}; i,j=1,...K

h_{ij} = the elements of the inverse G^{-1}

$\underline{g}(x)$ = the K-dimensional column vector whose i-th entry is $g(x-y_i;\sigma)$.

Consider the vector

$$\underline{g}^{\perp}(x) = G^{-1}\underline{g}(x), \qquad (2.6.2)$$

whose i-th entry is given by

$$g_i^{\perp}(x) = \sum_{j=1}^{k} h_{ij}\, g(x-y_j;\sigma) \qquad (2.6.3)$$

Postmultiply $\underline{g}^{\perp}(x)$ by the transpose of the vector $\underline{g}(x)$ and integrate each element of the resulting matrix over x:

$$\int \underline{g}^{\perp}(x)\, \underline{g}'(x)\, dx = G^{-1} \int \underline{g}(x)\underline{g}'(x)dx = I. \qquad (2.6.4)$$

I is the identity matrix. Hence, $g_i^{\perp}(x)$ is orthogonal to the space spanned by the $g(x-y_j;\sigma)$, j=1,...i-1,i+1,...K:

$$\int g_i^{\perp}(x)\, g(x-y_k;\sigma)dx = \sum_{j=1}^{K} h_{ij} \int g(x-y_j;\sigma)g(x-y_k;\sigma)dx$$

$$= 1 \text{ if } k = i \qquad (2.6.5)$$

$$= 0 \text{ otherwise.}$$

As an estimate of p_i at the n-th stage take

$$\hat{p}_{i,n} = \frac{1}{n} \sum_{\ell=1}^{n} g_i^{\perp}(X_\ell)$$

$$= \frac{1}{n} \sum_{j=1}^{K} h_{ij} \sum_{\ell=1}^{n} g(X_\ell - y_j;\sigma).^{1} \qquad (2.6.6)$$

The estimate is unbiased:

$$E\, \hat{p}_{i,n} = E\, g_i^{\perp}(X)$$

$$= \int g_i^{\perp}(x) f(x)dx$$

$$= \sum_{j=1}^{K} p_j \int g_i^{\perp}(x) g(x-y_j;\sigma)dx \qquad (2.6.7)$$

$$= p_i \,.$$

Observe that $g_i^{\perp}(x)$ is a bounded function

$$|g_i^{\perp}(x)| \le \frac{1}{\sqrt{2\pi}\,\sigma} \sum_j |h_{ij}| = \frac{c_8(i)}{\sqrt{2\pi}\,\sigma} \,. \qquad (2.6.8)$$

Consequently, our previous theory is immediately applicable. By a proof

---

[1] Observe that the "active" part of the estimate is the inner sum. The h_{ij} are constants computed before the estimating procedure begins.

which is essentially identical to Theorem 2.4.1, we obtain

Lemma 2.6.1: Under the hypotheses of Lemma 2.4.1, the sequences of estimates $\hat{p}_{i,n}$, i=1,...K, converge in mean-square to p_i with the variance of the estimate dominated by

$$E\left\{(p_i-\hat{p}_{i,n})^2\right\} = V(\hat{p}_{i,n}) \leq \frac{2}{n}\,\frac{c_8^2(i)}{2\pi\sigma^2}\left(1+\frac{B_2}{1-\rho_*}+2(M-1)\right) . \tag{2.6.9}$$

We have identified the random variable Z as $Z(\omega_j)=y_j$. That is, the underlying probability space consists of the points $\omega=\{\omega_0,\omega_1\ldots\omega_n\ldots\}$, with $P(\omega_j) = p_j$. $Z_\ell(\omega_j)$ is then interpreted as saying that in the ℓ-th interval, y_j was transmitted. The M-dependence assumption represents transmission with a finite memory; the probability of transmitting y_j in the ℓ-th interval and y_i in the m-th interval is $p_{ij}^{m-\ell}$, which need not equal p_ip_j if $|m-\ell| < M$.

The extension to transmission with "infinite" memory comes essentially from Theorem 2.2.2. For example, we can require that

$$p_{ij}^{m-\ell} = p_{ij}^{\tau} \longrightarrow p_i\,p_j \quad \text{as } \tau\to\infty ,$$

and

$$\sum_{\tau=1}^{n} |(p_{ij}^{\tau}-p_ip_j)| \leq B_3\,n^{\delta}, \quad 0\leq\delta<1 . \tag{2.6.10}$$

Then, the variance of the estimate is bounded by

$$E\left\{(p_i - \hat{p}_{in})^2\right\} = V(\hat{p}_{i,n}) \leq \frac{2c_8(i)}{n\, 2\pi\, \sigma^2}\left[(1 + \frac{3B_2}{1-\rho_*}) + K^2 B_3 n^{\delta}\right]. \tag{2.6.11}$$

Convergence takes place at the rate $1/n^{1-\delta}$.[1]

There remains to show that the $g(x-y_i;\sigma)$ are linearly independent. Assume they are dependent. Then, with $a_i \neq 0$ for all i, the dependence assumption gives

$$\sum_{i=1}^{K} a_i g(x-y_i;\sigma) = 0 . \tag{2.6.12}$$

Take the Fourier transform of (12) and divide out the common $e^{-\frac{1}{2}\sigma^2 v^2}$ term:

$$\sum_{i=1}^{K} a_i\, e^{jvy_i} = 0.$$

Multiply through by e^{-jvy_k}. Using the mean-value property

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T e^{jvx}\, du = 1 \quad \text{if } x = 0$$

$$= 0 \quad \text{otherwise}$$

and the fact that the y_i are distinct, we get that $a_i = 0$, i=1,2...K. This is a contradiction. Hence, the $g(x-y_i;\sigma)$ are linearly independent.

The above procedure is clearly applicable to the case

---

[1] The K^2 factor comes from bounding the term $\tilde{D}_{T,2}$, i.e., summing (10) over all i and j. See (2.2.41).

$$\alpha(z) = \sum_{i=1}^{K} p_i \int u(z-y)d\beta_i(y) \qquad (1.4.3)$$

where the $\beta_i(y)$ are known distributions and the set p_i is taken as unknown. The density function of the random variable X is

$$f(x) = \sum_{i=1}^{K} p_i f_i(x)$$

where $$f_i(x) = \int g(x-y;\sigma)d\beta_i(y) \ .$$

Since the $f_i(x)$ are bounded functions, so will the corresponding $f_i^{\perp}(x)$ be bounded. Then, we need only require that the characteristic functions of the distributions $\beta_i(y)$ be linearly independent.

2.7 SUMMARY AND GENERALIZATIONS

In this chapter we have primarily been concerned with the problem of estimating a special, but not unimportant, univariate density function

$$f(x) = \int g(x-z;\sigma)d\alpha(z) \ .$$

Given the sequence of dependent random variables X=N+Z, we have displayed consistent estimates of f(x) and have obtained bounds on the rate of convergence. We have given two methods of estimating f(x) and another method of estimating the product s(x)f(x).

To apply the kernel method, one must recompute the contribution of all past observations at each stage in order to obtain an asymptotically unbiased estimate. Using a gaussian kernel, we have shown that this method

gives a rate of convergence equal to $O(1/n^{4/5})$.[1] The estimate is a density function in that it is non-negative and integrates to one. No knowledge of the underlying process is needed to form the estimate or to specify the rate of convergence.

To use the first series method, we require a moment condition to be able to specify a rate of convergence. No knowledge of the gaussian process is needed to form the estimate. The eigenfunction representation for $s(x)f(x)$, on the otherhand, does require knowing the standard deviation σ. Both series methods may lead to estimates which, at some point of the sequence, are negative over a finite range of x.

In practice, either series can be truncated with the remaining finite number of coefficients estimated recursively. We have already pointed out the dependence of the asymptotic error on the number of terms used.

We have also shown (Corollaries 2.4.1 and 2.5.1) that both series methods converge in the manner required to guarantee the convergence of the empirical Bayes procedure discussed in section 1.2.

Somewhat secondary to our purpose, but worthy of mention, is the fact that the L_2 representation can be applied to the more general problem of density estimation given a sequence independent observations. Suppose the density function $p(x)$ and its first three derivatives exist. Further, assume that $x^3p(x), x^2p'(x), xp''(x)$ and $p(x)'''$ are L_2 functions. Then, using the technique in section 2.4, we can estimate $p(x)$ with a mean-square

---

[1] This assumes appropriate conditions on the dependencies of the observation.

error = $O(1/n^{1/3})$. This does not compare favorably to the rate $O(1/n^{4/5})$ obtained by the kernel method which requires a minimal amount of assumptions on $p(x)$. However, in estimating a k-variate density function, we will see that the series method, with the same type of assumptions as in the univariate case, keeps the $1/n^{1/3}$ rate, while the kernel method leads to a rate which is considerably slower.

With the exception of section 2.5, the results we have obtained are not unique to the gaussian noise assumption, but to a class of processes, of which the gaussian in the most prominent member. Specifically, the technique we have used is applicable to any stationary bivariate density function which can be expanded in the form[1]

$$p_2(x,y) = p_a(x)p_b(y) \sum_{i,j} a_{ij}\, \theta_i^{(a)}(x)\, \theta_j^{(b)}(y) \tag{2.7.1}$$

with $$a_{ij} = \int p_2(x,y)\, \theta_i^{(a)}(x)\, \theta_j^{(b)}(y)dx\, dy \tag{2.7.2}$$

and where p_a and p_b are the marginal density functions of $p_2(x,y)$. The $\theta_i^{(a)}(x)$ are polynomials orthogonal with respect to the weight $p_a(x)$. The Mehler formula is a special case of this expansion (with convergence already established) wherein,

$$\begin{aligned} a_{ij} &= 0, \quad i \neq j \\ &= \frac{\rho^j}{j!}, \quad i = j \end{aligned} \tag{2.7.3}$$

[1] Equation (1) is called the Barrett-Lampard expansion [3], and has found other uses in noise theory [7,22,23].

$$\theta_j(x) = He_j(x) \tag{2.7.3}$$

That is, for the bivariate gaussian density, the expansion is diagonal and $p_a(x) = p_b(x)$.

It is this expansion which gave the desired cancellation of the product of the univariate gaussian densities and the first term of the series (2.2.11). Any bivariate density function which can be expressed as in (1) will give the desired cancellation, for it is easy to see that, in general, $\theta_o(x)= 1$ and $a_{oo}=1$. Hence, the first term of the expansion is just the product of the univariate densities. Then, in analogy to the development in section 2.2, one can obtain the bound

$$\iint |p_\tau(x,y) - p_a(x)p_b(y)|dxdy \leq \sum_{i,j} |a_{ij}(\tau)|, \tag{2.7.4}$$

where i and j are not both equal to zero. To dominate the variance expression corresponding to (2.2.8), the summability of

$$\frac{1}{n}\sum_{\tau=1}^{n}\sum_{i,j} |a_{ij}(\tau)| \tag{2.7.5}$$

would be investigated. The interpretation of summability in terms of the underlying noise process would now have to be made with reference to all the moments of the bivariate density and not just the correlation function as in the gaussian case.

In the Barrett-Lampard paper, the authors give another example of a bivariate density which admits a diagonal expansion and for which the

coefficients form a geometric progression. This is the case of narrow-band gaussian noise subjected to an instantaneous square law envelope detector. The expansion takes the form([3], Eq. 80):

$$p_\tau(x_1,x_2) = p(x_1)p(x_2) \sum_{j=0}^{\infty} (\mu^2(\tau))^j L_j(x_1/2\sigma^2)L_j(x_2/2\sigma^2), \qquad (2.7.6)$$

where x_i is the square of the envelope, $x_i=R_i^2$, and the density function of R is given by the Rayleigh density. The polynomials in this expansion are the Laguerre polynomials which are orthogonal on $0 \leq x < \infty$ with respect to the weight

$$p(x) = \frac{1}{2\sigma^2} \exp\left(\frac{-x}{2\sigma^2}\right) \qquad (2.7.7)$$

The quantity μ is defined as([3], Eq. 72)

$$\mu^2(\tau) = \frac{1}{\sigma^4} \int_0^\infty \int_0^\infty s_n(f_1)\, s_n(f_2) \cos 2\pi(f_1-f_2)\tau df_1\, df_2$$

where $s_n(f)$ is the power spectrum of the narrow-band gaussian noise and

$$\sigma^2 = \int_0^\infty s_n(f)df \ .$$

Assuming that $\mu(\tau) < 1$ for $\tau \neq 0$, (4) reduces to

$$\frac{1}{n} \sum_{\tau=1}^{n} \frac{\mu^2(\tau)}{1-\mu^2(\tau)} \ , \qquad (2.7.8)$$

in direct analogy to the development in section 2.2.

For the class of stationary Markov processes, Barrett and Lampard show that if the bivariate density admits a diagonal expansion then the

correlation function is an exponential function of time and the expansion takes the form

$$p(x_1,x_2)=p(x_1)p(x_2)\sum_{j=0}^{\infty} e^{-\lambda_j \tau}\theta_j(x_1)\theta_j(x_2) \quad . \qquad (2.7.9)$$

τ is the distance (in time) between the random variables x_1 and x_2. Wong and Thomas [47] have further characterized the Markov processes which admit the above expansion. This class is composed of three distinct types, all of whose univariate density functions belong to the Pearson system of distributions.[1] This class consists of:

a) the gaussian density with the associated Hermite polynomials.

b) the density $p(x) = \frac{1}{\Gamma(\alpha+1)} x^{\alpha} e^{-x}$ and the associated Laguerre polynomials. For $\alpha=n-\frac{1}{2}$, $n=0,1,2,\ldots$ we have the chi-square distribution, and for $\alpha=0$ we have the case just discussed ((7)).

c) the density function $p(x) = \frac{1}{2^{\alpha+\beta+1}} \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)}(1-x)^{\alpha}(1+x)^{\beta}$ which represents the Pearson type I system. This includes the uniform density and the density function of a sine wave with unit amplitude and random phase. The associated polynomials are the Jacobi polynomials which include, among others, the Legendre and Chebyshev polynomials.

With the marginal density function of the Markov sequence given by one of the above, the corresponding bivariate density function is given

---

[1] Their technique is to reduce the Fokker-Plank equation to a Sturm-Liouville eigenvalue problem. Necessary and sufficient conditions are then found under which the eigenfunctions form a complete set of orthogonal polynomials. These conditions include a differential equation which the univariate density must satisfy and which characterizes the Pearson system.

by (9).

In the context of our problem, these results can be applied not only to the noise, but to the $\{Z_\ell\}$ sequence as well. For suppose that the density function of $\alpha_\tau(z_1,z_2)$ is given by (9). Then, the quantity $\tilde{D}_{\tau,2}$ (Eq. (2.2.41)) can be written as:

$$\begin{aligned}\tilde{D}_{\tau,2} &= \iint G(x-z_1;\sigma)G(x-z_2;\sigma)\left[\alpha'_\tau(z_1,z_2)-\alpha'(z_1)\alpha'(z_2)\right]dz_1\,dz_2 \\ &\leq \iint \left|\alpha'_\tau(z_1,z_2)-\alpha'(z_1)\alpha'(z_2)\right|dz_1\,dz_2 \\ &\leq \sum_{j=1}^{\infty} e^{-\lambda_j \tau} \quad .\end{aligned}$$

Consequently, a theorem analogous to Theorem 2.2.2 can be obtained with a rate of convergence determined by the growth or summability of the quantity

$$\frac{2}{n^2}\sum_{\tau=1}^{n}(n-\tau)\sum_{j=1}^{\infty}e^{-\lambda_j\tau} \quad .$$

To extend the technique of section 2.5 to other noise processes, one must solve an eigenfunction problem where the kernel is the corresponding univariate density function of the noise. We suspect, but have not proved, that progress in this direction can be made for those processes whose univariate densities are the weights associated with the classical polynomials. These polynomials ([12], page 164) are just those polynomials mentioned earlier whose corresponding weights belong to the Pearson system. Most of the specific properties of the Hermite polynomials which we used

are common to the classical polynomials. In addition, those polynomials which are solutions to a Sturm-Liouville problem form a complete set and behave (in an interval) as does the usual Fourier series (Hobson [18], page 771).

CHAPTER 3

ESTIMATING THE DENSITY FUNCTION OF THE OBSERVATIONS—k-VARIATE CASE

3.1 INTRODUCTION

In this chapter our previous results are extended to an arbitrary k-variate density function. For the univariate case, convergence statements generally followed from the inequality

$$\iint |g_2(x_1,x_2;\sigma,\rho_T) - g(x_1;\sigma)g(x_2;\sigma)|dx_1dx_2 \leq \frac{|\rho_T|}{1-|\rho_T|} . \qquad (3.1.1)$$

This was derived in section 2.2, equation (2.2.15). In this chapter we will be concerned with dominating the integral

$$\iint |g_{2k}(\underline{x}_1,\underline{x}_2;M_T)-g_k(\underline{x}_1;A)g_k(\underline{x}_2;A)|d\underline{x}_1\, d\underline{x}_2 , \qquad (3.1.2)$$

where g_k and g_{2k} denote the k-variate and 2k-variate gaussian density functions. x_1 and x_2 are k dimensional vectors. x_1 represents the k samples from the ℓ-th interval and x_2 the k samples from the m-th interval. Both vectors have a covariance matrix A which, from the stationarity assumption, is independent of the interval. The covariance matrix M_T of dimension 2k, is given by

$$M_T = E\left\{\begin{bmatrix} x_1 \\ \hline x_2 \end{bmatrix}[x_1' \,|\, x_2']\right\} \qquad (3.1.3)$$

$$= \left[\begin{array}{c|c} A & B_\tau \\ \hline B'_\tau & A \end{array}\right] , \qquad (3.1.3)$$

with $\tau = m-\ell$, and ' denoting the transpose.

In the next section we majorize the 2k-fold integral by displaying a transformation which converts (2) into k double integrals of the form (1). This change of variables allows us to go directly to the Mehler formula without any further generalization of the Hermite polynomials introduced earlier. The majorant will now be a function of the eigenvalues of a certain matrix and not simply expressed in terms of the correlation coefficients. A rate of convergence can still be determined by investigating the properties of the autocorrelation function.

Having majorized (2), the extension of our previous results will be obvious. For this reason we simply state the results in section 3.3, commenting when there is a significant difference in the technique or final result.

3.2 DOMINATING THE 2k-FOLD INTEGRAL

Define the 2k-dimensional vector u as

$$u = \left\{\begin{array}{c} x_1 \\ \hline x_2 \end{array}\right\} , \qquad (3.2.1)$$

and the matrix N (of dimension 2k) by

$$N = \left[\begin{array}{c|c} A & 0 \\ \hline 0 & A \end{array}\right] . \qquad (3.2.2)$$

Equation (3.1.2) is written as

$$\int |g_{2k}(u;M_T) - g_{2k}(u;N)|\,du\ , \tag{3.2.3}$$

with du representing the 2k differentials $du_1 \ldots du_{2k}$.

As already mentioned, we want to show that a change of variables $u=T^{-1}v$ reduces (3) to k double integrals of the form (3.1.1). What is the same thing is to show that a non-singular transformation T takes M_T into

$$T'M_TT = \left[\begin{array}{c|c} C & R_T \\ \hline R_T' & C \end{array}\right] \tag{3.2.4}$$

and N into

$$T'M_TT = \left[\begin{array}{c|c} C & 0 \\ \hline 0 & C \end{array}\right] \tag{3.2.5}$$

Here, C and R are kxk diagonal matrices.

This simulatneous transformation is accomplished in two stages. The first step is to reduce N and M to diagonal matrices. Let T_1 be the transformation such that

$$T_1'\ M_T\ T_1 = \Lambda_T = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_{2k} \end{bmatrix} \tag{3.2.6}$$

and

$$T_1'\ N\ T_1 = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array}\right] \ . \tag{3.2.7}$$

This is the usual "double-diagonalization" procedure.

The second step is the reverse of this process. That is, we then show that there is an orthogonal transformation T_2 such that

$$T_2' \Lambda_\tau T_2 = \left[\begin{array}{c|c} C & R_\tau \\ \hline R_\tau' & C \end{array}\right] = \left[\begin{array}{c|c} I & R_\tau \\ \hline R_\tau' & I \end{array}\right]. \qquad (3.2.8)$$

Since T_2 is orthogonal (7) remains invariant under this transformation. The required transformation is then $T=T_1T_2$.

From (6), it is not immediately evident that Λ_τ (with 2k diagonal elements) is similar to (8) which has k free parameters (the diagonal elements of R). In addition, to apply Mehler's formula we will need the elements of R_τ strictly less than one.

We will establish that Λ_τ and (8) are indeed similar, and that the elements of R are less than one. The only assumption which we will make is that M_τ be a positive definite covariance matrix. By this we shall always mean strictly positive definite.

What we are concerned with here is a generalized characteristic-value problem, or what Gantmacher calls the pencil of quadratic forms. We will use some results from Gantmacher [14], Chapter 10, section 6.

To avoid a proliferation of subscripts we drop the τ subscript for the present and since we are consistently dealing with vectors we will not use any special notation for them.

Two real symmetric quadratic forms $M(x,x)=x'Mx$ and $N(x,x)=x'Nx$ determine the pencil of forms $M(x,x)-\lambda N(x,x)$. λ is a parameter.

Definition: If the form $N(x,x)$ is positive definite, the pencil $M(x,x)-\lambda N(x,x)$ is called regular.

Definition: The equation $|M-\lambda N| = 0$ is called the characteristic equation of the pencil of forms $M(x,x)-\lambda N(x,x)$.

From (3), with M p.d. (positive definite), A is also p.d. since it is a principal minor of M. Hence, N is p.d., and the pencil of forms is regular.

Theorem 3.2.1 (Gantmacher, page 310): The characteristic equation of a regular pencil of forms, $|M-\lambda N| = 0$, always has 2k real roots λ_j with the corresponding principal vectors $z^j=(z_{1j},z_{2j},\ldots,z_{2k,j})$, $j=1,\ldots 2k$, which satisfy

$$Mz^j = \lambda_j N z^j \quad , \quad j=1,2,\ldots 2k \; . \tag{3.2.9}$$

These principal vectors can be chosen such that the relations

$$N(z^i,z^j) = z^{i'} N z^j = \delta_{ij}, \; i,j=1,\ldots 2k \tag{3.2.10}$$

are satisfied. From (10), it follows that the z^j, $j=1,\ldots 2k$ are linearly independent.

The existence of the required transformation T_1 is assured by

Theorem 3.2.2 (Gantmacher, p. 314): If $Z = \{z_{ij}\}_1^{2k}$ is a principal matrix of a regular pencil of forms $M(x,x)-\lambda N(x,x)$, then the transformation

$$x = Zy \tag{3.2.11}$$

reduces the forms $M(x,x)$ and $N(x,x)$ simultaneously to sums of squares

$$\sum_{j=1}^{2k} \lambda_j y_j^2 \quad , \quad \sum_{j=1}^{2k} y_j^2 \quad , \tag{3.2.12}$$

where $\lambda_1,\lambda_2,\ldots,\lambda_{2k}$ are the characteristic values of the pencil $M(x,x) - \lambda N(x,x)$ corresponding to the columns $z^1,z^2,\ldots,z^{2k}$ of Z.

Conversely, if some transformation $(x=Zy)$ simultaneously reduces $M(x,x)$ and $N(x,x)$ to the above form, then Z is a principal matrix of the regular pencil of forms $M(x,x)-\lambda N(x,x)$.

The first theorem is proved by writing (9) as $N^{-1}Mz = \lambda_j z^j$ and then showing that $N^{-1}M$ is similar to some symmetric matrix for which the characteristic values are known to be real, etc... The second theorem is the usual statement of the double-diagonalization process.

As a consequence of Theorem 3.2.1, we know that the elements of Λ in (6) are real. A further characterization is obtained by noting that for any characteristic value and vector, we have

$$z^{j'} M z^j = \lambda_j z^{j'} N z^j \quad , \quad j=1,2,\ldots 2k. \tag{3.2.13}$$

Since M and N are both p.d., it follows that

$$\lambda_j > 0 \quad ; \quad j = 1,2,\ldots 2k \ . \tag{3.2.14}$$

We now want to show that k of the λ_j are determined by the remaining k characteristic values.

Let z_1 and z_2 represent k-dimensional vectors and write $z = \{-\frac{z_1}{z_2}-\}$, where z is a principal vector of (9). Expressing M and N in terms of A

and B, we obtain from (9) the following two equations:

$$(1-\lambda)Az_1 + Bz_2 = 0 \qquad (3.2.15)$$

$$B'z_1 + (1-\lambda)Az_2 = 0 \qquad (3.2.16)$$

From (15) we have

$$(1-\lambda)z_1 = -A^{-1}Bz_2 \ . \qquad (3.2.17)$$

Multiply (16) by $(1-\lambda)$ and use (17) to substitute for $(1-\lambda)B'z_1$. Equation (16) becomes

$$\{(1-\lambda)^2A-B'A^{-1}B\}z_2 = 0 \ . \qquad (3.2.18)$$

In an analogous manner we obtain an equation for the z_1 vector,

$$\{(1-\lambda)^2A-BA^{-1}B'\}z_1 = 0. \qquad (3.2.19)$$

The 2k characteristic values λ_j must satisfy $|M-\lambda_jN| = 0$. They must also satisfy the characteristic equations associated with (18) and (19). It is not difficult to show that the characteristic equations of the above two equations((18) and (19)) are equal.[1] Therefore, we need only consider one of them. Let

$$r = (1-\lambda) \qquad (3.2.20)$$

Equation (19) is written as

---

[1]Bellman, R., Introduction to Matrix Analysis, McGraw-Hill Book Co., 1960, p. 94.

$$\{r^2A-BA^{-1}B'\}z_1 = 0. \tag{3.2.21}$$

Observe that $BA^{-1}B'$ is symmetric. Since A is p.d., (21) is a regular pencil with the parameter r^2. From Theorem 3.2.1, we have that there are k real roots r_j^2, j=1,2,...k, with the corresponding linearly independent principal vectors z_1^j, j=1,...k.

The characteristic equation corresponding to (21) is a polynomial of degree k in r^2. Viewed as a polynomial of degree 2k in r, the 2k roots are equal to $\pm\sqrt{r_j^2}$, j=1,...k. From the definition or r, we have

$$\begin{aligned} \lambda_j &= 1 + \sqrt{r_j^2} = 1 + r_j \\ \lambda_{j+k} &= 1 - \sqrt{r_j^2} = 1 - r_j \quad , \quad j = 1,2,...k. \end{aligned} \tag{3.2.22}$$

Since λ_j is greater than zero ((14)), we have the bound

$$|r_j| < 1, \quad j=1,...k. \tag{3.2.23}$$

Using these results, (6) can now be written as

$$T_1MT_1 = \Lambda = \begin{bmatrix} 1+r_1 & & & & & \\ & 1+r_2 & & 0 & & \\ & & \ddots & & & \\ & & & 1+r_k & & \\ & & & & 1-r_1 & \\ & 0 & & & & \ddots \\ & & & & & & 1-r_k \end{bmatrix} \tag{3.2.24}$$

with the k values of r^2 given by the roots of the polynomial

$$|A^{-1}BA^{-1}B' - r^2I| = 0 \qquad (3.2.25)$$

As an aside, we remark that the 2k principal vectors of (9) are given by

$$z^j = \frac{1}{\sqrt{2}}\begin{Bmatrix} z_1^j \\ \hline z_2^j \end{Bmatrix}, \quad z^{j+k} = \frac{1}{\sqrt{2}}\begin{Bmatrix} z_1^j \\ \hline -z_2^j \end{Bmatrix} \quad j=1,2,\ldots k,$$

where z_1^j and z_2^j are the principal vectors (of dimension k) of (19) and (18). It is a simple matter to see that the 2k vectors thus defined are linearly independent. It is only slightly more difficult to show that they perform the double-diagonalization, i.e., (12) is satisfied.

We now show that Λ, as given by (24), and the matrix defined in (8) are similar. For this it suffices to show that their characteristic equations are the same.

From (24) we have,

$$|\Lambda - \gamma I| = \prod_{i=1}^{k} (1-r_i^2 - 2\gamma + \gamma^2) . \qquad (3.2.26)$$

To evaluate the characteristic equation of (8), we need to specify the diagonal matrix R. Choose the k positive values $r_j = +\sqrt{r_j^2}$, j=1,...k, and take R as

$$R = \begin{bmatrix} r_1 & & & \\ & r_2 & & \\ & & \ddots & \\ & & & r_k \end{bmatrix} . \qquad (3.2.27)$$

The characteristic polynomial of (8) is given by

$$\left| \begin{bmatrix} I & R \\ R & I \end{bmatrix} - \gamma \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \right| \tag{3.2.28}$$

To determine this characteristic polynomial, we use the same argument as in going from (15) to (19). We write

$$\begin{bmatrix} I & R \\ R & I \end{bmatrix} \begin{Bmatrix} w_1 \\ w_2 \end{Bmatrix} = \gamma \begin{Bmatrix} w_1 \\ w_2 \end{Bmatrix}$$

and obtain the characteristic equation for the vector w_1 (or w_2),

$$\{(1-\gamma)^2 I - R^2\} w_1 = 0 . \tag{3.2.29}$$

The characteristic equation in given by $|(1-\gamma)^2 I - R^2|$, which is identical to (26). Hence, the existence of the orthogonal transformation T_2 is assured.

In the context of our problem, and to summarize the results to date, we have

Theorem 3.2.3: We are given the 2k-order integral

$$\int_u |g_{2k}(u;M) - g_{2k}(u;N)| du \tag{3.2.3}$$

with $M = \begin{bmatrix} A & B_T \\ B_T' & A \end{bmatrix}$, and $N = \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}$.

If M is a positive definite covariance matrix, there exists a change of variables

$$u = T^{-1}v = (T_1T_2)^{-1}v \tag{3.2.30}$$

with $|T| \neq 0$ which takes (3) into

$$\int\!\!\cdots\!\!\int \Big| \prod_{i=1}^{k} g_2(v_i, v_{i+k}; 1, r_i) - \prod_{i=1}^{k} g(v_i;1)g(v_i;1) \Big| dv_1 \ldots dv_{2k} \tag{3.2.31}$$

The v_j are scaler variables and the k values of r_i^2 are given by the roots of

$$0 = |A^{-1}BA^{-1}B' - r^2I| \ . \tag{3.2.25}$$

Since $|r_k| < 1$, i=1,...k, we can use Mehler's formula to majorize (31). Let the symbol $\underline{j} \neq 0$ mean "excluding the term $j_1=j_2=\ldots j_k=0$." Then, using the expansion for the bivariate gaussian density, we have:

$$\int\!\!\cdots\!\!\int \Big| \prod_{i=1}^{k} g_2(v_i, v_{i+}k; 1, r_i) - \prod_{i=1}^{k} g(v_i;1)g(v_i;1) \Big| dv_1 \ldots dv_{2k}$$

$$= \int\!\!\cdots\!\!\int \Big| \prod_{i=1}^{k} g(v_i;1)g(v_{i+k};1) \sum_{j_i=0}^{\infty} \frac{r_i^{j_i}}{j_i!} \, He_{j_i}(v_i) \, He_{j_i}(v_{i+k}) \Big| \ .$$

$$dv_1 \ldots dv_{2k}, \ \underline{j} \neq 0 \ . \tag{3.2.33}$$

Bring the integrations inside the summation and bound the integrals as in (2.2.14). We then obtain

Corollary 3.2.1: Let M_T be a positive definite covariance matrix. It then follows that

$$\int_u |g_{2k}(u;M_T) - g_{2k}(u;N)| du \ \leq$$

$$\prod_{i=1}^{k} \sum_{j_i=0}^{\infty} r_i^{j_i} \quad , \quad \underline{j} \neq 0$$

$$= -1 + \prod_{i=1}^{k} \frac{1}{1-r_{i,\tau}} \quad , \tag{3.2.34}$$

where r_i^2, satisfies

$$|A^{-1}B_\tau A^{-1}B'_\tau - r^2_{i,\tau} I| = 0, \; i=1,2,\ldots k.$$

For k=1 we have A=1, $B_\tau=\rho_\tau$, $r^2_{1,\tau}=\rho^2_\tau$, and (33) reduces to (2.2.15).

The majorant in (34) will enter into the variance calculations summed over τ. We want to again interpret this sum in terms of the autocorrelation function $R(\tau)$. In analogy to section 2.2, designate

$$\sum_{\tau=1}^{n} D'_\tau = \sum_{\tau=1}^{n} \left\{ -1 + \prod_{i=1}^{k} \frac{1}{1-r_{i,\tau}} \right\}$$

$$= \sum_{\tau=1}^{n} \left\{ \frac{1 - \prod_{i=1}^{k} (1-r_{i,\tau})}{\prod_{i=1}^{k} (1-r_{i,\tau})} \right\} . \tag{3.2.35}$$

Let $r_*=\min_{\tau>1} (1-r_{i,\tau})$. Since $|r_{i,\tau}| < 1$, $r_* > 0$. Then, (35) is dominated by

$$\sum_{\tau=1}^{n} D'_\tau \leq \frac{1}{r_*} \sum_{\tau=1}^{n} \left\{ 1 - \prod_{i=1}^{k} (1-r_{i,\tau}) \right\} . \tag{3.2.36}$$

Writing out this expression gives

$$\sum_{\tau=1}^{n} D'_\tau \leq \frac{1}{r_*} \sum_{\tau=1}^{n} \left\{ \sum_{i=1}^{k} r_{i,\tau} - \sum_{\substack{i,j \\ i\neq j}}^{k} r_{i,\tau} r_{j,\tau} + \ldots + (-1)^k r_{1,\tau} r_{2,\tau} \ldots r_{k,\tau} \right\} \tag{3.2.37}$$

The j-th inner summation gives $k!/(j!(k-j)!)$ terms, where $j=1,2,\ldots k$. There are then 2^k-1 terms which are summed over τ. Since $|r_{i,\tau}| < 1$, for all i and τ, any term which contains a $r_{i,\tau}$ as a factor is dominated by the single term $r_{i,\tau}$ appearing in the first inner sum. Consequently, (37) can be bounded by

$$\sum_{\tau=1}^{n} D'_\tau \leq \frac{1}{r_*} \sum_{\tau=1}^{n} \frac{(2^k-1)}{k} \sum_{i=1}^{k} r_{i,\tau} . \qquad (3.2.38)$$

From the Schwarz inequality we have

$$\left(\sum_{i=1}^{k} r_{i,\tau} \right)^2 \leq k \sum_{i=1}^{k} r_{i,\tau}^2$$

and it follows that

$$\sum_{\tau=1}^{n} D'_\tau \leq \frac{(2^k-1)}{k^{1/2} r_*} \left(\sum_{i=1}^{k} r_{i,\tau}^2 \right)^{1/2} .$$

The quantity $\sum r_{i,\tau}^2$ is just the trace of the matrix $(A^{-1}BA^{-1}B')$. Hence

$$\sum_{\tau=1}^{n} D'_\tau \leq \frac{(2^k-1)}{k^{1/2} r_*} \sum_{\tau=1}^{n} \left\{ \text{trace } (A^{-1}B_\tau A^{-1}B'_\tau) \right\}^{1/2} . \qquad (3.2.39)$$

The trace is the sum of k^4 terms. Each of these terms involves the product of the correlation function evaluated at different arguments. The matrix A^{-1}, which also involves the autocorrelation function, does not depend on τ but only on the manner in which samples are taken in a particular

interval. Define:

$$\begin{aligned} a_{ij}^{-1} &= \text{elements of } A^{-1} \\ R_{ij}(\tau) &= \text{elements of } B_\tau,\ i,j=1,2,\ldots k \\ R_*(\tau) &= \max_{i,j} |R_{ij}(\tau)| \\ A(k) &= \sum_i \sum_j |a_{ij}^{-1}| \ . \end{aligned} \tag{3.2.40}$$

The trace is given by

$$\text{trace } (A^{-1}B\ A^{-1}B') = \sum_{j_1,j_2,j_3,j_4}^{k} a_{j_1j_3}^{-1}\ a_{j_2j_4}^{-1}\ R_{j_3j_2}(\tau)\ R_{j_1j_4}(\tau) \ .$$

Using the above definitions, we obtain the bound

$$\text{trace } (A^{-1}B\ A^{-1}B') \leq R_*^2(\tau)A^2(k)$$

and (39) becomes

$$\begin{aligned} \sum_{\tau=1}^{n} D_\tau' &\leq \frac{(2^k-1)}{k^{1/2}r_*}\ A(k) \sum_{\tau=1}^{n} |R_*(\tau)| \\ &= B_\delta(k) \sum_{\tau=1}^{n} |R_*(\tau)| \ . \end{aligned} \tag{3.2.41}$$

We are now in a position to use the results of section 2.2 to obtain

Corollary 3.2.2: If $R(\tau)$ satisfies condition A, from (2.2.21), we obtain

$$\sum_{\tau=1}^{n} D_\tau' \leq B_\delta(k)\ \sigma^2\left(B_1 + \frac{n^{1-\delta}}{(1-\delta)}\right). \tag{3.2.42}$$

If condition B is satisfied, then, from (2.2.23), we have

$$\sum_{\tau=1}^{n} D'_{\tau} \leq B_8(k)\, \sigma^2\, B_2 \ . \qquad (3.2.43)$$

With the exception of when B_2 is given by the Euler-Maclaurin summation formula, the constants B_1 and B_2 have the same meaning as in section 2.2. To use the Euler-Maclaurin formula, rather than sum $R_*(\tau)$ over τ, we first sum $R_j(\tau)$ over j, j=1,...k, for each τ. That is, we sum all elements of the first row of B_τ for each τ, and then sum over τ. B_2 is now given by (cf.(2.2.25)).

$$B_2 = \frac{1}{\sigma^2} \int_0^{\infty} (k|R(t)| + |R'(t)|)dt \ .$$

3.3 ESTIMATING THE k-VARIATE DENSITY FUNCTION

Rather than introduce a new set of constants we will use the same notation as in chapter 2 for those constants which play similar roles.

3.3a The Empirical Distribution Function

The k-variate empirical distribution function is given by[1]

$$F_n(\underline{x}) = \frac{1}{n} \sum_{\ell=1}^{n} U_\ell(\underline{X}_\ell) \ . \qquad (3.3.1)$$

The random variable $U_\ell(\underline{X}_\ell)$ is defined as

[1] The notation $\underline{y}$ is used to designate a vector.

$$U_\ell(\underline{X}_\ell) = 1 \quad \text{if} \quad \underline{X}_\ell \leq \underline{x} \tag{3.3.2}$$

$$= 0 \qquad \text{otherwise}$$

$\underline{X}_\ell$ is the sample vector in the ℓ-th interval, so that by $\underline{X}_\ell \leq \underline{x}$ we mean the set of inequalities $X_{\ell i} \leq x_i$, $i=1,2,\ldots k$, $\ell=1,2,\ldots n$. In analogy to Theorem 2.2.2., we have

Theorem 3.3.1: Given the sequence of identically distributed stationary random vectors $\underline{X}_i = \underline{N}_i + \underline{Z}_i$, with the k-variate density function

$$f(\underline{x}) = \int g_k(\underline{x}-\underline{z};A)\, d\alpha(\underline{z}). \tag{3.3.3}$$

Define:

$$\tilde{D}_{\tau,2} = \iint G_k(\underline{y}_1-\underline{z}_1;A)\, G_k(\underline{y}_2-\underline{z}_2;A) \Big[d\alpha_\tau(\underline{z}_1,\underline{z}_2) - d\alpha(\underline{z}_1)\, d\alpha(\underline{z}_2) \Big]. \tag{3.3.4}$$

where

$$G_k(\underline{y}-\underline{z};A) = \int_{-\infty}^{\underline{y}} g_k(\underline{x}-\underline{z};A)\, d\underline{x}$$

and

$$\alpha_\tau(\underline{z}_1,\underline{z}_2) = \Pr\left\{ \underline{Z}_\ell \leq \underline{z}_1 \ , \ \underline{Z}_{\ell+\tau} \leq \underline{z}_2 \right\}.$$

Assume

i) $R(t)$ satisfies condition B

ii) $\tilde{D}_{\tau,2}$ satisfies

$$\sum_{\tau=1}^{\infty} \tilde{D}_{\tau,2} = B_3 < \infty \ .$$

Then, the empirical distribution function is a consistent estimate with the variance bounded by

$$E\left\{\left[(F(\underline{x})-F_n(\underline{x}))^2\right]\right\} = V(F_n(\underline{x}))$$

$$\leq \frac{1}{n}\left[1 + 6\,\sigma^2\, B_8(k)\, B_2 + B_3\right] \ . \qquad (3.3.5)$$

3.3b The Orthogonal Representation

The k-variate density function (3) is expanded in the series[1]

$$f(\underline{x}) = f(x_1,x_2,\ldots x_k) = \sum_{j1}\cdots\sum_{jk} a_{j_1\ldots j_k}\, \varphi_j(x_1/\sigma_1)\ldots\varphi_{j_k}(x_k/\sigma_k) \ . \qquad (3.3.6)$$

The φ's are the one-dimensional Hermite functions as in (2.4.3). We will write (6) as

$$f(\underline{x}) = \sum_{j} \underline{a}_j\, \varphi_j(\underline{x}) \ . \qquad (3.3.7)$$

The estimate of $\underline{a}_j$ at the n-th stage is denoted by $\hat{\underline{a}}_{jn}$,

$$\hat{\underline{a}}_{jn} = \frac{1}{n}\sum_{\ell=1}^{n} \varphi_j(\underline{X}_\ell) \ . \qquad (3.3.8)$$

These estimates are unbiased. Since the function $\varphi_j(\underline{x})$ is bounded (see

[1] $f(\underline{x})$ is bounded by $[(2\pi)^k|A|]^{-1/2}$ and therefore it is L_2 in R_k.

(2.4.7)) by

$$|\varphi_j(\underline{x})| \leq c_1^k/(\pi^{k/4}(\sigma_1\sigma_2\ldots\sigma_k)^{1/2}) = c_2 , \qquad (3.3.9)$$

in analogy to Theorem 2.4.1 and using Corollary 3.2.2, we have

Lemma 3.3.1: Assume the sequence of vectors $\{\underline{Z}_\ell\}$ is M-dependent and that $R(t)$ satisfies condition B. Then,

$$E\left\{(\underline{a}_j-\hat{\underline{a}}_{jn})^2\right\} = V(\hat{\underline{a}}_{jn}) \leq \frac{2\,c_2^2}{n}(1+\sigma^2 B_8(k)\,B_2+2(M-1)) = \frac{c_3}{n} . \qquad (3.3.10)$$

For $k = 1$, $\sigma^2 B_8(k)\,B_2 = \frac{2\,B_2}{1-\rho_*}$, and (10) reduces to (2.4.10).

As an estimate of $f(\underline{x})$, we take

$$\hat{f}_n(\underline{x}) = \sum_{j_1=0}^{q_1(n)}\cdots\sum_{j_k=0}^{q_k(n)} \hat{a}_{j_1\ldots j_k}\,\varphi_{j_1}(x_1/\sigma_1)\ldots\varphi_{j_k}(x_k/\sigma_k) .$$

We shall set $q_1 = \ldots = q_k = q$, and write the estimate as

$$\hat{f}_n(\underline{x}) = \sum_{j=0}^{q(n)} \hat{\underline{a}}_{jn}\,\varphi_j(\underline{x}) . \qquad (3.3.11)$$

Assume the r-th absolute product moment of $\underline{Z}$ exists:

$$\int\cdots\int |z_1z_2\ldots z_k|^r\, d\alpha(z_1,z_2,\ldots z_k) < \infty,\ r \geq 2 . \qquad (3.3.12)$$

Then, Lemma 2.4.3 can be directly extended to k dimensions. This results in the bound

$$\underline{a}_j^2 = a^2_{j_1\ldots j_k} \leq B_6/(j_1\ldots j_k)^r , \qquad (3.3.13)$$

and we have

Theorem 3.3.2: The MISE of the estimate is

$$J_n = E \int (f(\underline{x}) - \hat{f}_n(\underline{x}))^2 d\underline{x} = \sum_{j_1=q+1}^{\infty} \cdots \sum_{j_k=q+1}^{\infty} a^2_{j_1 \cdots j_k}$$

$$+ \sum_{j_1=0}^{q} \cdots \sum_{j_k=0}^{q} E(\hat{a}_{j_1 \cdots j_k} - a_{j_1 \cdots j_k})^2 \quad .$$

With (12) and Lemma 3.3.1 holding, we obtain the bound

$$J_n \leq B_\theta \left(\frac{1}{(r-1)} \frac{1}{q^{r-1}} + \frac{1}{q^r}\right)^k + \frac{q^k c_3}{n} \quad . \tag{3.3.14}$$

Choosing $q(n)$ as the largest integer $\leq (B_\theta n/c_3)^{\frac{1}{rk}}$, the MISE satisfies

$$J_n = O(1/n^{\frac{r-1}{r}}). \tag{3.3.15}$$

This is the same rate as in the univariate case. Naturally, the actual bound is different. The constant B_θ in (14) is now the L_2 norm of a function defined in R_k. This function is

$$(g(x_1;\sigma_1)\ldots g(x_k;\sigma_k))^{-1} \sum^{r} \frac{r!}{j_1!j_2!\ldots j_k!} (\sigma_1^{2j_1}\ldots\sigma_k^{2j_k})$$

$$\frac{d^r}{dx_1^{j_1}\ldots dx_k^{j_k}} [f(x_1\ldots x_k)g(x_1;\sigma_1)\ldots g(x_k;\sigma_k)] \quad ,$$

where the summation is to be taken over all possible integers with $\sum_{i=1}^{k} j_i = r$.

To investigate the mean-square error, define the function

$$f_q(\underline{x}) = \sum_{j_1=0}^{q} \cdots \sum_{j_k=0}^{q} a_{j_1 \ldots j_k} \varphi_j(x_1/\sigma_1) \ldots \varphi_j(x_k/\sigma_k) \ . \tag{3.3.16}$$

As in the one-dimensional case, this function converges to $f(\underline{x})$ pointwise and uniformly in any finite interval. In particular, with $r \geq 3$, we have the bound

$$|f(\underline{x}) - f_q(\underline{x})| < c_2 \sqrt{B_6} \left(\frac{1}{(\frac{r}{2}-1)\, q^{\frac{r}{2}-1}} + \frac{1}{q^{\frac{r}{2}-1}} \right)^k . \tag{3.3.17}$$

In analogy to Corollary 2 4.1, we obtain

Corollary 3.3.1: Assuming that Lemma 3.3.1 holds and that (12) is true with $r \geq 3$, we have

$$E_n \left\{ \left[\hat{f}_n(\underline{X}_n) - f(\underline{X}_n) \right]^2 \right\}$$

$$\leq \left\{ \left[c_2 \sqrt{B_6} \left(\frac{1}{(\frac{r}{2}-1)\, q^{\frac{r}{2}-1}} + \frac{1}{q \cdot \frac{r}{2}} \right)^k + c_2 \, q^k \sqrt{\frac{c_3}{n}} \right]^2 \right\} . \tag{3.3.18}$$

Upon choosing $q \sim 1/n^{1/rk}$, as $n \to \infty$, the mean-square error satisfies

$$E_n \left\{ \left[\hat{f}_n(\underline{X}_n) - f(\underline{X}_n) \right]^2 \right\} = O(1/n^{\frac{r-2}{r}}) \tag{3.3.19}$$

Observe that this is the same rate as in the univariate case.

3.3c The Eigenfunction Representation

We shall assume that the covariance matrix of the noise A is known and that the observation vector is "pre-whitened," i.e., we apply the transformation $A^{-1/2}$ to the observation vector $\underline{X}_\ell$. The new data vector is $A^{-1/2}(\underline{N}_\ell + \underline{Z}_\ell)$.[1] What this does is to make the resulting gaussian noise samples within an interval independent. The noise vectors in different intervals are still correlated with a cross-correlation matrix equal to $A^{-1/2} B_T A^{-1/2}$. Note that the characteristic polynomial for r^2 remains the same. Hence, Corollary 3.2.2 is applicable without change.

We assume that the data is pre-whitened without changing notation. The density function of the observation vector is

$$f(x_1,\ldots x_k) = \int\cdots\int g(x_1-z_1;1)\ldots g(x_k-z_k;1)d\alpha(z_1\ldots z_k) \qquad (3.3.20)$$

which we write as

$$f(\underline{x}) = \int_{\underline{z}} g_k(\underline{x}-\underline{z};I)d\alpha(\underline{z}) \ . \qquad (3.3.21)$$

With $\varphi_j(\underline{z})$ given in terms of the Hermite functions

$$\varphi_j(\underline{z}) = \varphi_{j_1}(z_1/\sigma_1)\ \varphi_{j_2}(z_2/\sigma_1)\ \ldots\ \varphi_{j_k}(z_k/\sigma_1) \ , \qquad (3.3.22)$$

where $\sigma_1 > 1$, define

---

[1] All that is necessary is to transform the data vector so that the resulting covariance matrix is diagonal. For purposes of notation, it is more convenient to have the resulting covariance matrix equal to the identity matrix. Hence we use $A^{-1/2}$. Since A is (strictly) positive definite, $A^{-1/2}$ is well defined.

$$\underline{d}_j = d_{j_1 \ldots j_k} = \int_{\underline{z}} \varphi_j(\underline{z}) d\alpha(\underline{z}) \ . \tag{3.3.23}$$

The coefficient $\underline{d}_j$ is uniformly bounded in j and $\underline{z}$,

$$|\underline{d}_j| = |d_{j_1 \ldots j_k}| \leq \left(\frac{c_1}{\pi^{1/4}\sigma} \right)^k \ . \tag{3.3.24}$$

Let,

$$s(\underline{x}) = \left(\frac{\sigma_1^2+1}{\sigma_1^2-1} \right)^{k/4} \exp\left\{ -\frac{1}{2} \frac{1}{(\sigma_1^2+1)(\sigma_1^2-1)} \sum_{i=1}^{k} x_i^2 \right\} \tag{3.3.25}$$

and define

$$\begin{aligned} \xi &= \frac{\sigma_1^2-1}{\sigma_1^2+1} \\ \gamma^2 &= \frac{(\sigma_1^2-1)(\sigma_1^2+1)}{\sigma_1^2} \end{aligned} \tag{3.3.26}$$

In analogy to section 2.5, we obtain the expansion

$$\begin{aligned} s(\underline{x})f(\underline{x}) &= \sum_{j_1=0}^{\infty} \cdots \sum_{j_k=0}^{\infty} (\xi^{j_1} \ldots \xi^{j_k}) d_{j_1 \ldots j_k} \varphi_{j_1}(x_1/\gamma) \ldots \varphi_{j_k}(x_k/\gamma) \\ &= \sum_{j} \underline{d}_j \ \underline{\xi}^{j} \varphi_j(\underline{x}) \ . \end{aligned} \tag{3.3.27}$$

This series converges in mean-square and uniformly in $\underline{x}$. We also have the uniform bound

$$\begin{aligned} |s(\underline{x})f(\underline{x}) - \sum^{q} \underline{d}_j \ \underline{\xi}^{j} \varphi_j(\underline{x})| &\leq \left(\frac{c_1^2}{\sqrt{\pi\sigma_1 \gamma}} \frac{\xi^{q+1}}{(1-\xi)} \right)^k \\ &= \left(\frac{c_1^2}{\sqrt{\pi}} \frac{\xi^{q+1}}{(1-\xi)} \frac{1}{(\sigma_1^4-1)} \right)^k \ . \end{aligned} \tag{3.3.28}$$

The unbiased estimates of the coefficients are

$$\underline{\xi}^{j}\,\hat{\underline{d}}_{jn} = \frac{1}{n}\sum_{\ell=1}^{n}\varphi_{j}(\underline{X}_{\ell})s(\underline{X}_{\ell}) \tag{3.3.29}$$

Using Corollary 3.2.2, we have

Lemma 3.3.2: Assume that $R(t)$ satisfies condition B and that the sequence $\{\underline{Z}_{\ell}\}$ is M-dependent. Then, the variance of the estimate is dominated by

$$V(\underline{\xi}^{j}\underline{d}_{j}) \leq \frac{2\,c_4^{\,2}}{n}\left\{1 + B_8(k)\sigma^{2}B_2 + 2(M-1)\right\} = c_5/n\ . \tag{3.3.30}$$

where c_4 is defined by

$$|\varphi_{j}(\underline{x})s(\underline{x})| \leq \left\{\frac{c_1}{\pi^{1/4}\sqrt{\gamma}}\left[\frac{\sigma_1^{2}+1}{\sigma_1^{2}-1}\right]^{1/4}\right\}^{k}$$

$$= \frac{c_1^{\,k}}{\pi^{k/4}}\ \frac{1}{(\gamma\xi)^{k/2}} = c_4 \tag{3.3.31}$$

The estimate of $s(x)f(x)$ is taken as

$$s(\underline{x})\hat{f}_{n}(\underline{x}) = \sum_{j_1=0}^{q}\cdots\sum_{j_k=0}^{q}\xi^{(j_1+\ldots j_k)}\,\hat{d}_{j_1\ldots j_{k_1}}\,\varphi_{j}(x_1/\sigma_1)\ldots\varphi(x_k/$$

$$= \sum_{j=0}^{q}\underline{\xi}^{j}\,\hat{\underline{d}}_{jn}\,\varphi_{j}(\underline{x})\ . \tag{3.3.32}$$

In analogy to Theorem 2.5.2 and Corollary 2.5.1, we have the following results.

Theorem 3.3.3: Under the hypothesis of the previous lemma, the MISE is dominated by

$$J'_n = E\int(s(\underline{x})f(\underline{x}) - s(\underline{x})\hat{f}_n(\underline{x}))^2\,d\underline{x}$$

$$\leq \left(\frac{c_1^2}{\pi^{1/2}}\,\frac{\xi^2}{\sigma_1(1-\xi^2)}\,\xi^{2q}\right)^k + \frac{c_5 q^k}{n} \qquad (3.3.33)$$

Choosing $q(n)$ as the largest integer $\leq \frac{1}{2k}\,\frac{\ell n\ n}{|\ell n\ \xi|}$, we have

$$J'_n \leq \left(\frac{c_1^2}{\pi^{1/2}\sigma_1(1-\xi^2)}\right)^k \frac{1}{n} + \frac{c_5}{(|\ell n\xi|2k)^k}\,\frac{(\ell n\ n)^k}{n}\ , \qquad (3.3.34)$$

or

$$J'_n = O((\ell n\ n)^k/n).$$

Corollary 3.3.2: Under the hypothesis of the previous lemma, the mean-square error in the estimate of the random variable $s(\underline{X}_n)f(\underline{X}_n)$ is bounded by $E_n\left\{\left[(s(\underline{X}_n)^2(f(\underline{X}_n)-\hat{f}_n(\underline{X}_n))^2\right]\right\}$

$$\leq\left\{\left[\left(\frac{c_1^2\,\xi^{q+1}}{\sqrt{\pi}\,(\sigma_1^4-1)(1-\xi)}\right)^k + \left(\frac{c_1}{\pi^{1/4}\gamma}\right)^k\sqrt{\frac{c_5}{n}}\,q^k\right]\right\}^2 . \qquad (3.3.35)$$

Letting q be the largest integer $\leq c_7 + \frac{\ell n\ n}{2k|\ell n\xi|}$ gives

$$E_n(\ ) \leq\left\{\left[\left(\frac{c_1^2\,\xi^{c_7}}{\sqrt{\pi}\,(\sigma_1^4-1)(1-\xi)}\right)^k\frac{1}{\sqrt{n}} + \left(\frac{c_1}{\pi^{1/4}\gamma}\left[c_7 + \frac{1}{2k}\,\frac{\ell n\ n}{|\ell n\xi|}\right]\right)^k\cdot\sqrt{\frac{c_5}{n}}\right]\right\}^2, \qquad (3.3.36)$$

or,

$$E_n(\ ) = O((\ell n\ n)^{2k}/n)\ .$$

The constant c_7 is not given explicitly as in (2.5.33), but would be determined by minimizing (36).

3.3d The Gaussian Kernel

The estimate of f(x) is taken as

$$\hat{f}_n(\underline{x}) = \frac{1}{n} \sum_{\ell=1}^{n} g_k(\underline{x}-\underline{X}_\ell;H) \ . \tag{3.3.37}$$

With H a diagonal matrix with i-th entry equal to h_i^2, (37) is just the k-dimensional version of (C.1). We take H as this diagonal matrix.

The expectation of the estimate is

$$E\ \hat{f}_n(\underline{x}) = \int_z d\alpha(\underline{z}) g_k(\underline{x} - \underline{z};\ A + H) \tag{3.3.38}$$

As in section 2.3a, the bias is dominated by considering the expression $(E\hat{f}_n(\underline{x}) - f(\underline{x}))$ and expanding in a Taylor series. Since the kernel $g_k(\underline{y};H)$ is an even function in $\underline{y}$, the first and mixed second derivatives drop out. We obtain, as $n \to \infty$ and $h_i \to 0$, i=1,...k,

$$E\ \hat{f}_n(\underline{x})-f(x) \to \frac{1}{2}\,(2\pi)^{k/2} \sum_{i=1}^{k} \frac{\partial^2 f(\underline{x})}{\partial\, x_i^2}\, h_i^{\,2} + O(h_i^{\,4}). \tag{3.3.39}$$

The variance expression is

$$V(f_n(\underline{x})) = \frac{1}{n} \left\{ \left[E(g^2(\underline{x}-\underline{X};H) - \left[E(g(\underline{x}-\underline{X};H) \right]^2 \right\}$$

$$+ \frac{2}{n^2} \sum_{\ell=1}^{n} \sum_{m=\ell+1}^{n} \left[E(g(\underline{x}-X_\ell;H) g(\underline{x}-X_m;H)) \right.$$

$$-E(g(\underline{x}-\underline{X}_\ell;H)\,Eg(\underline{x}-\underline{X}_m;H)\Big]\Big\} \qquad (3.3.40)$$

Using (38), the first bracket is bounded by

$$\frac{1}{n}\left\{\left[(2\pi)^{2k}(h_1\ldots h_k)\left|A+\frac{H}{2}\right|\right]^{-1}+\left[(2\pi)^k|A+H|\right]^{-1}\right\}. \qquad (3.3.41)$$

For the second expression, the term inside the double summation is given by[1]

$$Q_{m-\ell}=\int_{\underline{z}_1}\int_{\underline{z}_2} d\alpha(\underline{z}_1)d\alpha(\underline{z}_2)\left[g_{2k}(\underline{x}-\underline{z}_1,\underline{x}-\underline{z}_2;M+\tilde{H})-\right.$$

$$\left.g_{2k}(\underline{x}-\underline{z}_1,\underline{x}-\underline{z}_2;N+\tilde{H})\right]. \qquad (3.3.42)$$

M and N have the same meaning as before, and $\tilde{H}=\left[\begin{array}{c|c} H & 0 \\ \hline 0 & H \end{array}\right]$. Since $M+\tilde{H}$ is positive definite, we can apply the transformation of the previous section to the expression in the bracket in (42). We then dominate the resulting expression using Mehler's formula and Cramer's bound (see(C.8) and (C.9)). The result is,

$$Q_{m-\ell} = Q_\tau \le \left(\frac{c_1^2}{2\pi}\right)^k\left[-1+\prod_{i=1}^{k}\frac{1}{1-r_{i,\tau}}\right]. \qquad (3.3.43)$$

The $r_{i,\tau}^2$ are the roots of $|(A+H)^{-1}B_\tau(A+H)^{-1}B_\tau' - r^2I|$. Use Corollary 3.2.2 to obtain

$$\frac{2}{n^2}\sum_{\ell}\sum_{m} Q_{m-\ell} \le \left(\frac{c_1^2}{2\pi}\right)^k \frac{\sigma^2 B_8(k)B_2}{n}.$$

---

[1] Here, we assume the sequence $\{\underline{Z}_\ell\}$ is independent.

The constant $B_8(k)$ is defined differently since $A(k)$ is now the bound on the elements of $(A+H)^{-1}$.

The variance is dominated by

$$V(\hat{f}_n(\underline{x})) \leq \frac{1}{n}\left[\left(\frac{c_1}{2\pi}\right)^{2k} B_8(k)\sigma^2 B_2 + \frac{1}{(2\pi)^k}\,\frac{1}{|A+H|}\right.$$

$$\left. + \frac{1}{(2\pi)^{2k}|A+\frac{H}{2}|}\,\frac{1}{(h_1\ldots h_k)}\right.$$

$$\left. = \frac{b_1}{n} + \frac{b_2}{n(h_1\ldots h_k)}\right] \qquad (3.3.44)$$

Adding the square of the bias error to (44), the mean-square error is

$$E\left\{(f(\underline{x}) - \hat{f}_n(\underline{x}))^2\right\}$$

$$\leq \frac{b_1}{n} + \frac{b_2}{n(h_1 h_2\ldots h_k)} + \sum_{i,j}^{k} b_{ij}\, h_i^2\, h_j^2 \qquad (3.3.45)$$

where the b_{ij} are constants which bound the second partials in (39).

We have not made much progress in selecting the set h_i so as to minimize (45). The obvious thing to do is to set $h_1=\ldots h_k=h$. Then, it is easy to see that h is chosen proportional to $1/n^{1/(k+4)}$ and

$$E\left\{(f_n(\underline{x})-f(\underline{x}))^2\right\} = O(1/n^{4/(4+k)}) \; . \qquad (3.3.46)$$

From the way in which we have taken the estimates, it is not surprising that the assumptions needed to specify a rate of convergence (for all three methods) are direct extensions of those needed for the univariate

case. Thus, throughout this chapter we have assumed that the covariance matrix M_τ is (strictly) positive definite. In Chapter 2, we required the same thing by taking $|\rho_\tau| < 1$, i.e., for k=1,

$$M_\tau = \begin{bmatrix} 1 & \rho_\tau \\ \rho_\tau & 1 \end{bmatrix}$$

and $|\rho_\tau| < 1$ implies that M_τ is positive definite.

To use the eigenfunction representation, we take the covariance matrix A as known. For the L_2 series, we require a product moment assumption in order to specify a rate. Again, no knowledge of the underlying process is required to form the kernel estimate or to specify a rate of convergence. The eigenfunction representation gives essentially the same 1/n rate as in the univariate case: for the MISE we have the rate $(\ln(n))^k/n$ and for the mean-square error we obtain $(\ln(n))^{2k}/n$. With a r-th product moment assumption on the vector $\underline{Z}$, we obtain the same rate of convergence for the L_2 series as in the univariate case. The kernel method now gives a slower rate—the reason being that the bias in the estimate is still $O(h^4)$ while the variance term is now $O(1/h^k)$.

CHAPTER 4

APPLICATIONS OF THE EMPIRICAL BAYES TECHNIQUE

4.1 INTRODUCTION

We now apply our results to some problems in communication theory. As discussed in Chapter 1, we will be concerned with procedures which converge to what we have called the optimum one-stage test. To reiterate, this test uses only the present observations for the present test and, as such, is truely optimum only when the sequence of observations is independent. The empirical sequence of tests makes decisions on the same basis as the one-stage procedure, but it incorporates all past observations in updating the estimate of the test function. Furthermore, we will take a sequence of tests, the test function of which is identical in structure to the test function one would use if all distributions were known. It does not follow, especially in the small sample case, that this is the optimum thing to do. What we expect to show is that when we are repeatedly faced with the same decision problem, our sequence of test functions will get closer, in the mean-square sense, to the one-stage test. It then will follow from the results of section 1.2 that the empirical risk will approach the risk incurred by using the one-stage procedure.

So far, we have considered estimating the marginal density function of the observation which, in general, can be written as

$f(x) = \int g(x-z;\sigma)d\alpha(z)$. What remains is to show that we can extract from the estimate of $f(x)$ those quantities needed to form a consistent estimate of the test function. In the supervisory mode, there is no difficulty in finding these consistent estimates. Since we have available samples which are correctly classified with probability one, we can estimate the particular density $f_j(x)$ from which it was drawn. As would be expected, obtaining consistent estimates of the test function is more difficult when operating under a nonsupervisory condition.

In the remainder of this section, we discuss the problem of transmission through a random unknown channel when learning samples are available. This problem relates the results of section 1.2 on the convergence of the empirical procedure with the results on density estimation. It also serves to indicate when we can expect to find a solution to the nonsupervisory problem.

In section 4.2 we consider the problem of transmitting known signals with unknown a priori probabilities.

The problem of communication through a random multiplicative channel is considered in section 4.3. This problem is discussed in some detail since, for the case of small nonlinear distortion, a first-order analysis reduces to an analysis of a multiplicative disturbance.[1]

---

[1] One of K signals, $y_i(t)$, is transmitted with gaussian noise added to a distorted version of the signal. The received waveform is $x(t) = n(t)+y_i(t,\tau)$, where τ is, say, a random delay with known mean value τ_o. If, for example, the variance of τ is small, we write $\tau=\tau_o+\Delta\tau$ and approximate $x(t)$ by $x(t)=n(t)+y_i(t,\tau_o)+\Delta\tau(\partial/\partial\tau)y_i(t,\tau)\Big|_{\tau=\tau_o}$. Then, $\Delta\tau$ is taken as a zero mean random variable with an unknown distribution.

A problem with an unbounded loss function is discussed in section 4.4.

4.1a Communication Through an Unknown Random Channel—Supervised Learning

Suppose we transmit one of two signals, $y_0(t)$ or $y_1(t)$, with a priori probabilities p_0 and $p_1 = 1-p_0$. The signal is passed through a random unknown channel. We take the output of the channel to be a stationary random process $z_0(t)$ or $z_1(t)$, depending on which signal was transmitted. The received waveform during any interval is then

$$x(t) = n(t) + z_i(t), \quad i=0,1, \quad \ell \leq t \leq (\ell+1). \tag{4.1.1}$$

The density function of a single time sample is

$$f(x) = p_0 \, f_0(x) + p_1 \, f_1(x) \tag{4.1.2}$$

where

$$f_0(x) = \int g(x-z_0;\sigma) d\alpha(z_0) \tag{4.1.3}$$

$$f_1(x) = \int g(x-z_1;\sigma) d\beta(z_1) \, . \tag{4.1.4}$$

The distributions α and β are taken as unknown and may or may not be related.[1]

The statistical inference problem is to decide, with minimum probability of error, which of the two processes $z_0(t)$ or $z_1(t)$ is present

[1] α and β would be unrelated if, instead of transmitting known signals through a random channel, the problem was one of sending one of two unrelated random signals to which gaussian noise is added.

during each interval. The one-stage procedure is to evaluate the test function

$$T(x) = p_1 f_1(x) - p_o f_o(x) \tag{4.1.5}$$

and compare it to a zero threshold.

Suppose that we have an estimate of $T(x)$ and, after we have made a decision, we are told whether or not the decision was correct. In this "supervised learning" situation, we can use the observation to update the estimate of the density function from which it was drawn. We now have a better estimate of the particular density and the test function which we subsequently use for the next decision. Assuming p_o is known, the error in the estimate of the test function after the n-th decision would be

$$\hat{T}_n(x) - T(x) = p_1(f_1(x)-\hat{f}_{1n_1}(x))-p_o(f_o(x)-\hat{f}_{on_o}(x)) . \tag{4.1.6}$$

n_i is the known (after a decision is made) number of occurrences of $y_i(t)$ in the n intervals, $n_o+n_1=n$. Let $\tilde{\gamma}_{in_j}=E_n\left\{(f_i(X)-f_{in_j}(X))^2\right\}$. Then, by the Minkowski inequality, we have

$$E_n\left\{(\hat{T}_n(X_n)-T(X_n))^2\right\} \leq (p_1 \tilde{\gamma}_{1n_1} + p_o \tilde{\gamma}_{on_o})^2 = \gamma_n^2 . \tag{4.1.7}$$

Hence, to guarantee that γ_n tends to zero, say at the same rate as for the case of independent samples, we need to require that the autocorrelation function satisfy condition B, e.g., it be integrable and

eventually monotonically decreasing. In addition, we need a weak dependency amongst the samples derived from $z_i(t)$ as reflected in Theorem 2.2.2. Furthermore, to guarantee that the probability of error P_{en} converges to P_e (the probability of error using the one-stage test), we also require that (see Theorem 1.2.2)

a) $f_i(x) \neq 0$ a.e. x, i=0,1,

and to be able to specify a bound on the rate of convergence of P_{en} it is sufficient to assume that

b) $f_i(x)$ is analytic, i=0,1

c) $f_i(x)$, i=0,1 are linearly independent, i.e., unequal.

From (3) and (4), we see that conditions a and b are satisfied as $g(x-z;\sigma)$ is non-negative and analytic. If α and β are unequal then condition c will be satisfied.

We will illustrate the estimation procedure with the eigenfunction representation. As in section 2.5, define:

$$d_j = \int_{-\infty}^{+\infty} \varphi_j(z/\sigma_1)d\alpha(z)$$

$$e_j = \int_{-\infty}^{+\infty} \varphi_j(z/\sigma_1)d\alpha(z)$$

$$\xi^2 = \frac{\sigma_1^2 - \sigma^2}{\sigma_1^2 + \sigma^2}$$

$$\gamma^2 = \frac{(\sigma_1^2-\sigma^2)(\sigma_1^2+\sigma^2)}{\sigma_1^2} = \frac{(\sigma_1^2+\sigma^2)^2}{\sigma_1^2}\ \xi^2 \quad . \tag{4.1.8}$$

Multiplying (2)-(4) by

$$s(x) = \frac{1}{\sqrt{\xi}}\ \exp - \frac{x^2}{2}\left(\frac{\sigma}{\sigma_1\gamma}\right)^2 \tag{4.1.9}$$

gives

$$s(x)f_o(x) = \sum_{j=0}^{\infty} \xi^j\, d_j\, \varphi_j(x/\gamma) \tag{4.1.10}$$

$$s(x)f_1(x) = \sum_{j=0}^{\infty} \xi^j\, e_j\, \varphi_j(x/\gamma) \tag{4.1.11}$$

$$s(x)f(x) = \sum_{j=0}^{\infty} \xi^j\, (p_o d_j + p_1 e_j)\varphi_j(x/\gamma). \tag{4.1.12}$$

A test function equivalent to (5) is given by

$$s(x)T(x) = \sum_{j=0}^{\infty} (p_1 e_j - p_o d_j)\xi^j\, \varphi_j(x/\gamma) \quad . \tag{4.1.13}$$

Operating under a supervisory condition, we make a decision and then we are told from which population X_n was drawn.[1] Assuming X_n was drawn from $f_1(x)$, we then update the estimates of e_j as in section 2.5. At the end of the n-th decision, our estimate of the test function would be

---

[1] Perhaps more appropriate to communication type problems, learning samples can also be provided by transmitting a known sequence y_o, y_1, y_o, y_1, etc... interspersed with the message sequence.

$$s(x)\hat{T}_n(x) = p_1 \sum_{j=0}^{q_1(n_1)} \hat{e}_{jn_1} \xi^j \varphi_j(x/\gamma) - p_o \sum_{j=0}^{q_o(n_o)} d_{jn_o} \xi^j \varphi_j(x/\gamma) \ . \qquad (4.1.14)$$

Supposing that each $q_i(n_i)$ is chosen as in Corollary 2.5.1, we obtain the bound

$$E_n\left\{s^2(X_n)\ (\hat{T}_n(X_n)-T(X_n))\right\}^2 = (\gamma_n')^2 = O(\ln^2 n_*/n_*) \ , \qquad (4.1.15)$$

where $n_* = \min(n_o,n_1)$. Hence, from Corollary 1.2.2, the difference between the probability of error of the empirical procedure and the one-stage procedure is bounded by

$$0 \le P_{en} - P_e \le \left(\frac{2(\gamma_n')^2}{\epsilon^2} + \delta'(\epsilon)\right) \qquad (4.1.16)$$

where $\delta'(\epsilon) = \Pr\left\{|s(x)T(x)| < \epsilon\right\}$ and $(\gamma_n')^2$ is given by (15). Therefore, with the above assumed dependencies on the observations, P_{en} converges at the rate $\ln^2 n_*/n_*$.

Recall that the constant σ_1 is chosen greater than σ but otherwise arbitrary. We would naturally choose σ_1 to minimize the difference between P_{en} and P_e. To illustrate how σ_1 enters into (16), suppose we truncate the series (14) at q+1 terms. Then, using the bound in (2.5.37) for $(\gamma_n')^2$, the asymptotic difference in the probabilities of error is dominated by

$$0 \leq \lim_{n\to\infty} P_{en}-P_e \leq \frac{2}{\epsilon^2}\left\{\frac{c_1^4}{\pi}\,\frac{1}{4\sigma^4}\,\frac{(\sigma_1^2-\sigma^2)^{2q+1}}{(\sigma_1^2+\sigma^2)^{2q+3}}\right\} + \delta'(\epsilon)\ . \tag{4.1.17}$$

The closer σ_1 is to σ, the smaller the first term of (17). On the other-hand, from (8) and (9) we have

$$s(x) = \sqrt{\frac{\sigma_1^2+\sigma^2}{\sigma_1^2-\sigma^2}}\ \exp\left(\frac{x^2}{2}\,\frac{\sigma^2}{(\sigma_1^4-\sigma^4)}\right)$$

and $\delta'(\epsilon)$ is seen to increase as σ_1 approaches σ.

One possible procedure to follow would be to choose σ_1 so that $\delta'(\epsilon)$ is less than some preassigned small number Δ. Then, q is chosen large enough so that the first term of (17) is also less than Δ.

When more than one sample is used to base a decision, the assumptions and results are direct extensions of those for the univariate case. Thus, if we took k samples for each decision, P_{en} would converge to (a smaller) P_e at the rate $0(\ln^{2k} n_*/n_*)$.

This discussion assumes that we operate in a supervisory mode and, if one is willing to use a one-stage (or finite-stage) test, the procedure just outlined is straightforward and provides a solution to a variety of problems.[1] However, the above formulation and solution are

[1] Included in this formulation is the case of non-coherent communication. By restricting attention to a one-stage test we do, however, exclude the case of intersymbol interference.

not the best we can do. The above procedure has eliminated the signal design problem—the known signals $y_0(t)$ and $y_1(t)$ do not influence the convergence of the procedure. This has come about as we have neglected the relationship between $\alpha(z_0)$ and $\beta(z_1)$. A relationship certainly exists as the channel presumably distorts both signals in the same manner.

That some relationship between $\alpha(z_0)$ and $\beta(z_1)$ must exist and be utilized in order to learn when operating in a nonsupervisory mode is almost obvious. For, with no knowledge of the relationship between α and β, there is no way to extract from $f(x)$ consistent estimates of $f_0(x)$ and $f_1(x)$. In contrast, if α and β are equal then f_0 and f_1 are also equal and there is no longer any statistical inference problem. It is somewhere between these two cases where solutions to the nonsupervisory problem are to be found.

One such case is the transmission of known signals through a random multiplicative channel. We consider this problem in section 4.3 where it will be seen that the nature of the signals enters into the bounds and, in fact, determines whether or not consistent procedures can be found in the nonsupervisory mode of operation. For the supervisory mode, the rate given above in (15) will be improved upon with n replacing n_*—the point being that, in some cases, both sets of coefficients, d_j and e_j, can be updated at each stage.

4.1b The Detection of Noise in Gaussian Noise

We want to mention a somewhat different application of the eigen-

function representation, the detection of noise in gaussian noise when all distributions, including the a priori probabilities, are known. The application has the unusual aspect of incorporating a test procedure which is sequential in the number of terms of the series used for the test.

For simplicity, we restrict our attention to decisions based on a single sample. Using the notation of the previous sub-section, assume one of two random processes, $z_0(t)$ or $z_1(t)$, is transmitted with a priori probabilities p_0 and p_1. With the distributions $\alpha(z_0)$ and $\beta(z_1)$ known then, in principal, the coefficients d_j and e_j are known. The procedure which minimizes the probability of an incorrect decision is to evaluate (13) and compare it to a zero threshold.

Consider truncating the test function at q+1 terms,

$$s(x)T_q(x) = \sum_{j=0}^{q+1} (p_1 e_j - p_0 d_j)\xi^j \varphi_j(x/\gamma) . \qquad (4.1.18)$$

From (2.5.13), we have the bound

$$\left| s(x)f_0(x) - \sum_{j=0}^{q} \xi^j d_j \varphi_j(x/\gamma) \right| \leq \frac{c_1^2}{\sqrt{\pi\sigma_1\gamma}} \frac{\xi^{q+1}}{1-\xi} \qquad (4.1.19)$$

with an identical bound for the truncated series for $s(x)f_1(x)$. Then, the difference between (13) and (18) is dominated by

$$|s(x)(T(x)-T_q(x))| \leq \frac{c_1^2}{\sqrt{\pi\sigma_1\gamma}} \frac{\xi^{q+1}}{1-\xi} = \frac{c_1^2}{\sqrt{\pi(\sigma_1+\sigma)}} \frac{\xi^{q-1/2}}{1-\xi} \quad (4.1.20)$$

Hence, if for a given observation X, the value of $s(x)T_q(x)$ is greater in magnitude than the right side of (20), the decision using the truncated test function is the same as when the complete series is used.

Again, we would not pick σ_1 arbitrarily close to σ so as to make ξ arbitrarily small. All this does is scale down the possible range of values of $s(x)T_q(x)$ and $s(x)T(x)$. Specifically, from (2.4.3), we have

$$\varphi_j(x/\gamma) = \frac{g(x/\gamma)H_j(x/\gamma)}{\sqrt{2^j j!/\sqrt{4\pi}}\,\gamma} \quad (2.4.3)$$

and $s(x)T_q(x)$ becomes

$$s(x)T_q(x) = \frac{1}{\sqrt{\gamma}\sqrt{\pi}} \exp\left(-\frac{x^2}{2}\frac{1}{\xi^2}\frac{\sigma_1^2}{(\sigma_1^2+\sigma^2)^2}\right)\sum_{j=0}^{q} \xi^j$$

$$(p_1 e_j - p_o d_j)\frac{H_j(x/\gamma)}{\sqrt{2^j j!}} \quad (4.1.21)$$

Having fixed the value of σ_1, the procedure would be to evaluate the first q_1 terms of (21) and compare the magnitude to (20). If the magnitude is greater than (20), we announce hypothesis H_1 if $s(x)T_{q_1}(x)$ is positive and H_o if it is negative. If $|s(x)T_{q_1}(x)|$ is less than (20), compute another term of the series and recycle. In this manner, we expect to eventually make a decision which would be identical to the

decision based on the original test function. The value of q at which the procedure terminates is a random variable whose distribution depends on the value of σ_1 chosen, the coefficients d_j and e_j, and the observation X.

There is no theoretical difficulty in extending the procedure to a finite number of samples. We use the nonsingular transformation $A^{-1/2}$ to whiten the gaussian noise. Then, for example, the vector $\underline{z}_o$ is transformed into $\underset{\sim}{z}_o$ with a distribution $\alpha(A^{1/2}\underline{z}_o)$. The difficulties, of course, are in inverting A and in calculating the coefficients $\underline{d}_j$.

4.2 MULTIPLE (SIMPLE) HYPOTHESES WITH UNKNOWN A PRIORI PROBABILITIES

We consider the problem of detecting one of K+1 known signals when the a priori probabilities are unknown. In each time interval, $\ell \leq t \leq \ell + 1$, $\ell = 0,1,\ldots$, a signal $y_j(t)$ is chosen with a priori probability p_j, $j=0,1,\ldots K$. Zero mean, correlated gaussian noise is added to the signal. The received waveform is then $x(t)=n(t)+y_j(t)$. In the next sub-section, we consider the detection problem with a finite number of time samples. In sub-section 4.2b, the Karhunen-Loève expansion is used to obtain limiting forms.

4.2a Finite Number of Observations Per Decision

Under the j-th hypothesis, the density function for k time samples $\underline{x}=(x_1,x_2,\ldots x_k)$ is given by

$$f_j(\underline{x}) = g_k(\underline{x}-\underline{y}_j;A) \qquad (4.2.1)$$

where A is the non-singular covariance matrix of the gaussian noise vector $\underline{n}$, and $\underline{y}_j$ represents the k samples of the signal $y_j(t)$. The overall density function is

$$f(\underline{x}) = \sum_{j=0}^{K} p_j f_j(\underline{x}). \tag{4.2.2}$$

For a minimum probability of error test based on k samples, as discussed in section 1.2, we form the (K+1) test functions

$$T_j(\underline{x}) = p_j f_j(\underline{x}) - p_o f_o(\underline{x}), \; j = 0,1,\ldots K. \tag{4.2.3}$$

The decision function is $t_j(\underline{x}) = 1$ if $T_j(\underline{x}) > T_i(\underline{x})$, $i = 0,1\ldots K$, and $t_j(\underline{x}) = 0$ otherwise.

Suppose that the set of a priori probabilities are unknown. Then in (3), we use estimates of these quantities which are updated in every transmission interval. In the n-th interval the error in estimating the test function is

$$\hat{T}_{nj}(\underline{x}_n) - T_j(\underline{x}_n) = (\hat{p}_{jn} - p_j) f_j(\underline{x}_n) - (\hat{p}_{on} - p_o) f_o(\underline{x}_n), j=0,1\ldots K. \tag{4.2.4}$$

$\underline{x}_n$ denotes the n-th sample vector $\underline{X}_n = \{X_{1n}, X_{2n} \ldots X_{kn}\}$. $\hat{p}_{jn}$ is naturally a function of all the observations $\{\underline{X}_\ell\}$, $\ell = 1,2,\ldots$.

Observe that $f_j(\underline{x})$ contains a common factor $(1/(2\pi)^{k/2}|A|^{1/2})$ which can be cancelled out of the K+1 test functions. We assume this has been done in (4), but keep the notation unchanged. Since $f_j(\underline{x})$

is a bounded function, for any unbiased estimate $\hat{p}_{jn}$, we have

$$E_n\left\{(\hat{p}_{jn}-p_j)^2 f_j^2(\underline{X}_n)\right\} \leq V(\hat{p}_{jn})/(2\pi)^{k/2}|A|^{1/2} \qquad (4.2.5)$$

The mean-square error of (the modified) equation (4) is dominated by

$$E_n\left\{(\hat{T}_{nj}(\underline{X}_n)-T_j(\underline{X}_n))^2\right\} \leq \left(\sqrt{V(\hat{p}_{jn})} + \sqrt{V(\hat{p}_{on})}\right)^2 = \gamma_{jn}^2, j=1,2,\ldots K. \qquad (4.2.6)$$

Notice that we begin the index at j=1 since we would naturally set $\hat{T}_{no}(\underline{X})=0$, which equals $T_o(\underline{X})$.

The inequality in (6) is the bound we need to dominate the difference in the probabilities of error. Let P_{en} be the probability of error using the empirical procedure and P_e the probability of error when the a priori probabilities are known. In analogy to the equation above (1.2.40) and to (1.2.46), which are valid for one sample per interval, we have for k samples per interval:

$$P_e = 1-p_o - \sum_{j=0}^{K} \int_{A_j} T_j(\underline{x})d\underline{x} \qquad (4.2.7)$$

$$A_j = A_j\left\{\underline{x}: T_j(\underline{x}) \geq T_i(\underline{x}),\ i=0,1,\ldots K\right\},$$

and

$$0 \leq P_{en}-P_e \leq \sum_{j=0}^{K}\left\{\frac{2(\gamma_{jn}+\gamma_{u(j)n})^2}{\epsilon_{ju(j)}^2} + \delta_{ju(j)}(\epsilon_{ju(j)})\right\} \qquad (4.2.8)$$

with γ_{on} set equal to zero. The subscript u(j) plays the role of k in

(1.2.46) since, in this section, k is the number of observations per interval. In addition, here we let u=u(j) depend on the index j. The reason for this is as follows. Recall that $\epsilon_{ju(j)}$ is an arbitrary constant and $\delta_{ju(j)} = \Pr\left\{\left[|T_j(\underline{x})-T_u(\underline{x})| < \epsilon_{ju}\right]\right\}$. One would like to choose the subscript u, u(j)=0,1,...K, so that $\delta_{ju(j)}$ is a minimum for each j. This involves solving for the roots of $T_j(\underline{x})-T_u(\underline{x})=0$ in terms of the signals and a priori probabilities and then, perhaps by linearization of T_j-T_u, obtaining bounds for the δ_{ju} in terms of the signals $\underline{y}_j$ and $\underline{y}_u$.

There remains to estimate the p_j's so as to bound the difference $P_{en}-P_e$. To do this, we proceed as in section 2.6. We use the sequence of observations x_{in} with i fixed (e.g., the first component of each observation vector) to obtain an unbiased estimate $\hat{p}_{jn}$. Assuming the samples of the signals are distinct, $y_{ij}\neq y_{ij_1}$, $j,j_1=0,1,...K$ and i=1,...k, there are k such sequences available which give k unbiased estimates of p_j. These are then combined in a linear fashion.

Taking the sequence of observations $x_{i\ell}$, $\ell=1,2,...n$, the density function of $x_{i\ell}$ is

$$f(x_{i\ell}) = \sum_{j=0}^{K} p_j g(x_{i\ell}-y_{ij\ell};\sigma) .$$

We assume the samples of the signals, y_{ij}, j=0,1,...K, are distinct. Then, the estimate of p_j is taken as (see (2.6.6)):

$$p_{jn} = \frac{1}{n} \sum_{\ell=0}^{n} g_j^{\perp}(X_{i\ell})$$

$$= \frac{1}{n} \sum_{v=1}^{K} h_{jv} \sum_{\ell=1}^{n} g(X_{i\ell}-y_{iv\ell};\sigma), \quad j=0,1,\ldots K.$$

Now, assume that the autocorrelation function of the gaussian noise is integrable and eventually monotonically decreasing (condition B). Assume further that the signal transmitted in a particular interval can depend only on the signals transmitted in the previous M-1 intervals (M-dependence). Then, the hypotheses of Corollary 2.6.1 hold and we obtain $V(\hat{p}_{jn})=0(1/n)$. Using this in (8), we have P_{en} converging to P_e at the rate 1/n. Convergence also takes place at the 1/n rate with an "infinite transmission memory" if the conditional a priori probabilities satisfy (2.6.10) with $\delta=0$.

The above procedure gives no guarantee that the estimates $\hat{p}_{jn}$ are probabilities. In practice, we would want to normalize the estimates so that $0 \leq \hat{p}_{jn} \leq 1$ and $\sum_{j=0}^{K} \hat{p}_{jn}=1$.

4.2b Limiting Forms

The application of the empirical Bayes procedure to the finite sample case is straightforward. This is not so when limiting forms are considered. To use the Karhunen-Loève expansion, we make the following definitions:[1]

[1]Since all processes are stationary, these definitions hold for any interval. For the properties of the expansion see [11].

$$\int_0^1 R(s-t)\ \psi_j(t)dt = \lambda_j\psi_j(s),\ 0 \le s \le 1, j=1,2,\ldots$$

$$x_j = \int_0^1 x(t)\psi_j(t)dt$$

$$n_j = \int_0^1 n(t)\psi_j(t)dt \qquad (4.2.9)$$

$$y_{ij} = \int_0^1 y_i(t)\psi_j(t)dt,\ i=0,1,\ldots K.$$

We then have as the first k observations

$$x_j = n_j + y_{ij},\ j=1,2,\ldots k\ ,$$

and it follows from a property of the expansion that

$$E(n_j n_i) = 0,\ i \neq j$$

$$\lambda_j,\ i = j\ .$$

Since the n_j are uncorrelated gaussian random variables, they are independent. Hence, under the i-th hypothesis, the density function of the first k observations $\underline{x} = (x_1 \ldots x_k)$ is simply the product of k univariate gaussian density functions

$$f_i(\underline{x}) = f_i(x_1 \ldots x_k)$$

$$= \frac{1}{(2\pi)^{k/2}(\lambda_1 \ldots \lambda_k)^{1/2}} \exp - \frac{1}{2}\left\{\sum_{j=1}^{k} \frac{(x_j - y_{ij})^2}{\lambda_j}\right\}. \qquad (4.2.10)$$

Define the functions

$$v_{ik}(t) = \sum_{k_1=1}^{k} \frac{y_{ik_1}\,\psi_{k_1}(t)}{\lambda_{k_1}} \quad , \; i=0,1,\ldots K; \; 0 \le t \le 1 \tag{4.2.11}$$

and the inner products

$$(v_{ik},x) = \int_0^1 v_{ik}(t)\, x(t)dt = \sum_{k_1=1}^{k} \frac{y_{ik_1}x_{k_1}}{\lambda_{k_1}} \quad , \; i=0,\ldots K. \tag{4.2.12}$$

$$(v_{ik},y_i) = \int_0^1 v_{i_k}(t)\, y_i(t)dt = \sum_{k_1=1}^{k} \frac{y_{ik_1}^2}{\lambda_{k_1}} \tag{4.2.13}$$

Now, it is more convenient to take the test functions as the ratio

$$T_j(\underline{x}) = \frac{p_j}{p_o}\,\frac{f_j(\underline{x})}{f_o(\underline{x})} \quad , \quad j=0,\ldots K. \tag{4.2.14}$$

The decision function remains the same. Cancelling common factors in (14) and using the above definitions and (10) yields

$$T_j(\underline{x}) = \frac{p_j}{p_o} \exp -\frac{1}{2} \left\{ \left[-2(v_{jk},x)+(v_{jk},y_j) \right. \right.$$

$$\left. \left. +2(v_{ok},x)-(v_{ok},y_o) \right] \right\}.$$

Since the logarithm is a monotonic function we can just as well use $\ln T_j(\underline{x})$ in the test.

$$\ln T_j(\underline{x}) = \ln(p_j/p_o) + (v_{jk}-v_{ok},x) + \frac{(v_{ok},y_o)-(v_{jk},y_j)}{2} . \tag{4.2.15}$$

We shall assume

$$\sum_{k_1=1}^{\infty} \frac{y_{ik_1}^2}{\lambda_{k_1}^2} < \infty \quad , \quad i = 0,\ldots K. \tag{4.2.16}$$

Then, it is known ([32]) that:

i) the series in (12) converges with probability one

$$\lim_{k\to\infty} (v_{ik},x) = (v_i,x) = \sum_{k_1=1}^{\infty} \frac{y_{ik_1}x_{k_1}}{\lambda_{k_1}} < \infty, \; i = 0,\ldots K. \tag{4.2.17}$$

ii) $v_{ik}(t)$ converges to an L_2 function $v_i(t)$ which satisfies the integral equation

$$\int_0^1 R(t-s)v_i(s)ds = y_i(t), \quad 0 \le t \le 1, \; i = 0,\ldots K. \tag{4.2.18}$$

Since we are assuming that the autocorrelation function is strictly positive definite, the solution to (18) is unique.

Define the random variables

$$w_i = \int_0^1 x(t)v_i(t)dt, \qquad i = 0,\ldots K. \tag{4.2.19}$$

and the quantity

$$u_{ij} = (v_i,y_i) = \int_0^1 v_i(t)\, y_i(t)dt.^1 \tag{4.2.20}$$

The (gaussian) random variable w_i is the output of the filter

---

[1] u_{ij} can be thought of as the signal correlation to noise ratio. The white noise case gives: $R(t)=N\delta(t)$, $u_{ij} = \int_0^1 y_i(t)y_j(t)dt/N_0$, and u_{ii}= signal energy/noise power density (per cycle).

matched to the signal $y_i(t)$.

The logarithm of the test function is now a function of the random variables (w_j-w_o). Taking the limit as $k \to \infty$ in (15) yields:

$$\ln(T_j(w_j-w_o)) = \ln(p_j/p_o)+(w_j-w_o)+ \frac{u_{oo}-u_{jj}}{2} \qquad (4.2.21)$$

The error in the estimate of the log of the test function is simply

$$\ln(\hat{T}_{nj}(w_j-w_o)) - \ln(T_j(w_j-w_o))$$

$$= \ln(\hat{p}_{jn}/\hat{p}_{on}) - \ln(p_j/p_o)$$

$$= \ln(\hat{p}_{jn}/p_j) - \ln(\hat{p}_{on}/p_o) \quad . \qquad (4.2.22)$$

To form the estimate of p_j, consider the output of the r-th matched filter during any time interval $\ell \leq t \leq (\ell+1)$. Assuming the i-th hypothesis to be active during this interval, the received waveform is $x(t) = n(t)+y_i(t)$, and w_r is a gaussian random variable with mean value u_{ri} and variance equal to u_{rr}. Averaging over all possible hypotheses, the density function of the observation w_r is

$$f(w_r) = \sum_{i=0}^{K} p_i \; g(w_r-u_{ri}; \sqrt{u_{rr}}) \qquad (4.2.23)$$

The only difference between the situation here and in section 2.6 is that the correlation between the random variable (w_r-u_{ri}) in different intervals is not given directly in terms of $R(t)$. We do, however, have:

$$E\left\{\left[(w_{r_i\ell}-u_{r_i})(w_{r_im}-u_{r_i})\right]\right\} = \int_m^{m+1}\int_\ell^{\ell+1} E(n(t_1)n(t_2))v_r(t_1)v_r(t_2)dt_1dt_2$$

$$= \int_0^1\int_0^1 R(m-\ell+t_2-t_1)v_r(t_1)v_r(t_2)dt_1dt_2 \qquad (4.2.24)$$

If we assume that R(t) satisfies condition B ((2.2.23)), let $\tau=m-\ell$, and denote (24) by $u_{rr}\tilde{\rho}_\tau$, then,

$$\frac{1}{n}\sum |\rho_\tau| \le \frac{\sigma^2}{u_{rr}} B_2\left[\int_0^1 |v_r(t)|dt\right]^2$$

$$\le \frac{\sigma^2}{u_{rr}} B_2 b_r^2 , \qquad (4.2.25)$$

where b_r is a bound on the L_1 norm of $v_r(t)$. Since $\int_0^1 |v_r(t)|dt \le \left\{\int_0^1 v_r^2(t)dt\right\}^{1/2}$, we can set b_r equal to the L_2 norm of $v_r(t)$. From the definition of $v_r(t)$ and the previous assumptions, this norm is finite.

Using the output of the r-th matched filter, we take as the estimate of p_j

$$\hat{p}_{jn} = \frac{1}{n}\sum_{\ell=1}^{n} g_j(w_{r\ell})$$

$$= \frac{1}{n}\sum_{i=1}^{K} h_{ji}\sum_{\ell=1}^{n} g(w_{r\ell}-u_{ri};\sqrt{u_{rr}}),\ j = 1,\ldots K. \qquad (4.2.26)$$

$w_{r\ell}$ is the output of the r-th matched filter during the ℓ-th interval. As in section 2.6, these estimates are unbiased. In analogy to the previous case of time samples, assuming the autocorrelation function

satisfies condition B and that the signals satisfy an M-dependency, from Corollary 2.6.1 and (25), the variance expression is dominated by

$$V(\hat{p}_{jn}) \leq \frac{2}{n} \frac{c_s^2(j)}{2\pi\, u_{rr}} \left(1 + \frac{\sigma^2}{u_{rr}} B_2\, b_r^2 + 2(M-1)\right)$$

$$= \tilde{\gamma}_{jn}^2 \quad , \quad j = 1,\ldots K. \tag{4.2.27}$$

To calculate the mean-square error in the estimate of the test function (22), we proceed as follows. Since $\hat{p}_{jn}$ converges to p_j in mean-square, it converges in probability:

$$\Pr\left\{\left[|\hat{p}_{jn}-p_j| \geq \epsilon_j\right]\right\} \leq \tilde{\gamma}_{jn}^2/\epsilon_j^2 \ .$$

Consequently, with probability greater than $1-\tilde{\gamma}_{jn}^2/\epsilon_j^2$,

$$\ln(\hat{p}_{jn}/p_j) = \ln(1+(\hat{p}_{jn}-p_j)/p_j) = \frac{\hat{p}_{jn}-p_j}{p_j} + \frac{1}{p_j^2} O(\epsilon_j^2) \ .$$

Letting A equal the set of sample points which satisfy $\int_A dP(\omega) \geq 1-\tilde{\gamma}_{jn}^2/\epsilon_j^2$, it follows that

$$\int_A \ln^2(p_{jn}(\omega)/p_j)\,dP(\omega) \leq \frac{\tilde{\gamma}_{jn}^2}{p_j^2} + O(\epsilon_j^2) \tag{4.2.28}$$

Upon applying the Minkowski inequality to the expected value of the square of (22), it follows that, except for a set of experiments of probability less than $\tilde{\gamma}_{jn}/\epsilon_j^2$ or $\tilde{\gamma}_{on}^2/\epsilon_o^2$ (whichever is greater), the mean-square error in the j-th test function is dominated by

$$E_n\left\{\left[(\ln\hat{T}_{nj}-\ln T_j)^2\right]\right\} \leq \left(\gamma_{jn} + \frac{p_j^2+p_o^2}{p_j^2\ p_o^2}\ O(\epsilon_j)\right)^2 , \quad j=1,\ldots K. \tag{4.2.29}$$

We have defined $\gamma_{jn} = (\tilde{\gamma}_{jn}/p_j+\tilde{\gamma}_{on}/p_o)$ and as before, we set $\hat{T}_{no}(w_o-w_o)=0$ which gives $\gamma_{on}=0$. Letting $\gamma_n^* = \max_j \gamma_{jn}/\epsilon_j$, in analogy to (1.2.37), for this (equivalent) procedure we obtain: except in a set of experiments of probability less than γ_n^{*2}, the difference in the probabilities of error is bounded by

$$0 \leq P_{en}-P_e \leq \sum_{j=0}^{K} \left\{ \frac{2}{\epsilon_{jk}^2}\left\{\gamma_{jn} + \gamma_{kn} + \frac{p_j^2+p_o^2}{p_j^2\ p_o^2}\ O(\epsilon_j)+ \right.\right.$$

$$\left.\left. \frac{p_k^2+p_o^2}{p_k^2\ p_o^2}\ O(\epsilon_k)\right]^2 + \delta_{jk}(\epsilon_{jk})\right\}. \tag{4.2.30}$$

From our earlier assumptions, $\tilde{\gamma}_{jn}^2 = O(1/n)$. Hence, P_{en} converges to P_e at the same rate, except for a set of possible experiments of probability less than γ_n^{*2}.

To investigate the manner in which the signals affect the bounds, we consider the case of binary communication, K=1. The difference in the probabilities of error is then given by Theorem 1.2.2,

$$0 \leq P_{en}-P_e \leq \frac{2\ \gamma_n^2}{\epsilon^2} + \delta(\epsilon) , \tag{4.2.31}$$

where γ_n^2 is defined below (29) with j=1, $p_o+p_1=1$, and $\delta(\epsilon)$ is defined as $\delta(\epsilon)=\Pr\left\{\left|\ln T_1(w_1-w_o)\right| < \epsilon\right\}$.

Suppose we only use the output of the j=0 matched filter to estimate p_o. The bound on the variance of the estimate is

$$\tilde{\gamma}_{on}^2 \leq \frac{2}{n} \frac{c_B^2(0)}{2\pi u_{oo}} \left(1 + \frac{\sigma^2}{u_{oo}} \frac{B_2 b_r^2}{1-\rho_*} + 2(M-1)\right) \qquad (4.2.32)$$

where

$$c_B(0) = \sum_{j=0}^{1} |h_{oj}| .$$

The h_{ij} are the elements of the inverse of the Gramian matrix G (see (2.6.1)). For this example, G^{-1} is

$$G^{-1} = \frac{\begin{bmatrix} 1 & -\exp -(u_{oo}-u_{o1})^2/2u_{oo} \\ -\exp -(u_{oo}-u_{o1})^2/2u_{oo} & 1 \end{bmatrix}}{\sqrt{4\pi u_{oo}}\left(1-\exp\left\{-(u_{oo}-u_{o1})^2/u_{oo}\right\}\right)} ,$$

and $c_B(0)$ becomes

$$c_B(0) = \frac{1}{\sqrt{4\pi u_{oo}}} \frac{\left(1 + \exp\left\{-(u_{oo}-u_{o1})^2/2\,u_{oo}\right\}\right)}{\left(1 - \exp\left\{-(u_{oo}-u_{o1})^2/u_{oo}\right\}\right)}$$

The variance of the estimate is written

$$\tilde{\gamma}_{on}^2 = \frac{2}{n} \frac{1}{8\pi^2 u_{oo}^2} \left(\frac{\sigma^2}{u_{oo}} \frac{B_2 b_o^2}{1-\rho_*} + 2M-1\right) \left(\frac{1 + \exp\left\{-(u_{oo}-u_{o1})^2/2u_{oo}\right\}}{1 - \exp\left\{-(u_{oo}-u_{o1})^2/u_{oo}\right\}}\right) \qquad (4.2.33)$$

For purposes of illustration, let us assume that the autocorrelation

of the gaussian noise is approximately a delta function, $R(t)=N_o\delta(t)$. Then, $v_o(t)$ is approximately $y_o(t)/N_o$, and we have

$$b_o^2 \leq \int_0^1 v_o^2(t)dt \approx u_{oo}/N_o = u_{oo}/\sigma^2 ,$$

and,

$$\gamma_{on}^2 \approx \frac{2}{n} \frac{1}{8\pi^2 u_{oo}^2} \left(\frac{B_2}{1-\rho_*} + 2M-1\right) \left(\frac{1 + \exp\left\{-(u_{oo}-u_{o1})^2/2\, u_{oo}\right\}}{1 - \exp\left\{-(u_{oo}-u_{o1})^2/u_{oo}\right\}}\right) .$$

Roughly speaking, the variance of the estimate is inversely proportional to the square of the signal-to-noise ratio. Consequently, in combining estimates of p_o from the two matched filters we would weigh more heavily the filter corresponding to the larger signal energy.

For testing (K+1) hypotheses, there is a simplification when the signals are orthogonal. The signals $y_i(t)$ are said to be orthogonal if

$$u_{ij} = \int_0^1 v_i(t)y_i(t)dt = \int_0^1\int_0^1 v_i(t)R(t-s)v_j(s)dt\, ds = 0, \quad i\neq j .$$

For this situation, the density function of the output of the r-th matched filter (equation (23)) reduces to

$$f(w_r) = p_r\, g(w_r-u_{rr}; \sqrt{u_{rr}}) + (1-p_r)g(w_r; \sqrt{u_{rr}}) .$$

Hence, the output of this filter is used to estimate only p_r and operationally, the procedure for estimating the a priori probabilities

reduces to that given in the introductory example of section 1.1.

4.3 TRANSMISSION OF KNOWN SIGNALS THROUGH AN UNKNOWN RANDOM MULTIPLICATIVE CHANNEL

We now discuss the problem of detecting one of K+1 nonzero known signals which are passed through a random multiplicative channel. In any interval, the received waveform is

$$x(t) = N(t) + Z_\ell y_i(t); \quad \ell \leq t \leq (\ell+1), \quad i=0,\ldots K , \tag{4.3.1}$$

where the a priori probability of transmitting $y_i(t)$ is p_i. The signal is amplitude modulated by a random variable Z_ℓ, which may depend on the previous Z's, but which is independent of the gaussian noise. In this problem, we take the a priori probabilities, p_i, $i=0,1,\ldots K$, as known, and the a priori distribution of Z, $\alpha(z)$, as unknown. We will mainly be concerned with learning in the nonsupervised mode as the problem of supervised learning (with an arbitrary channel) has already been discussed in 4.1. We will, however, point out when the results in 4.1 can be improved upon for the particular case of a multiplicative channel.

At the end of this section, we shall briefly mention the problem where the received waveform is given by

$$x(t) = N(t) + Z(t) y_i(t); \quad \ell \leq t \leq (\ell+1), \quad i=0,1,\ldots K. \tag{4.3.2}$$

Here, the signals are amplitude modulated by the (unknown) random process Z(t).

We start with the problem given by (1) and again derive limiting forms via the Karhunen-Loève expansion. Using the notation and definitions of the previous section, from (4.2.9), the observations are the $X_{k_1}, k_1=1,\ldots k$. Under hypothesis H_j, the density function for the first k observations in the ℓ-th interval is

$$f_j(\underline{x}) = \int_{-\infty}^{+\infty} \frac{}{(2\pi)^{k/2}(\lambda_1\ldots\lambda_2)^{1/2}} \exp\left\{-\frac{1}{2}\sum_{k_1=1}^{k}\frac{(x_{k_1}-z_\ell\, y_{jk_1})^2}{\lambda_{k_1}}\right\} d\alpha(z_\ell)\,. \tag{4.3.3}$$

The density function of the observations averaged over all hypotheses is simply

$$f(\underline{x}) = \sum_{j=0}^{K} p_j f_j(\underline{x}). \tag{4.3.4}$$

For the (K+1) test functions we take

$$T_j(\underline{x}) = \frac{p_j\, f_j(\underline{x})}{p_o\, f_o(\underline{x})}\,, \qquad j = 0,1\ldots K. \tag{4.3.5}$$

Cancel common terms in (5) and use the definitions in (4.2.11)-(4.2.13) to obtain

$$T_j(\underline{x}) = \frac{p_j}{p_o}\,\frac{\displaystyle\int_{-\infty}^{+\infty}\left[\exp -\frac{1}{2}\left\{-2z_\ell(v_{jk},x)+z_\ell^2(v_{jk},y_j)\right\}\right] d\alpha(z_\ell)}{\displaystyle\int_{-\infty}^{+\infty}\left[\exp -\frac{1}{2}\left\{-2z_\ell(v_{ok},x)+z_\ell^2(v_{ok},y_o)\right\}\right] d\alpha(z_\ell)}\,. \tag{4.3.6}$$

Since $(v_{jk},y_j) = \sum_{k_1} \frac{y_{jk_1}}{\lambda_{k_1}} > 0$, $j=0,1,\ldots K$, we can complete the square in z_ℓ. Define the random variables

$$w_{jk} = \frac{(v_{jk},x)}{(v_{jk},y_j)} \quad , \qquad j = 0,1,\ldots K. \tag{4.3.7}$$

Then,

$$T_j(\underline{x}) = \frac{p_j}{p_o} \frac{e^{+\frac{w_{jk}^2}{2}(v_{jk},y_j)}}{e^{+\frac{w_{ok}^2}{2}(v_{ok},y_o)}} \frac{\int_{-\infty}^{+\infty}\left[\exp -\frac{1}{2}\left\{(v_{jk},y_j)(w_{jk}-z_\ell)^2\right\}\right]d\alpha(z_\ell)}{\int_{-\infty}^{+\infty}\left[\exp -\frac{1}{2}\left\{(v_{ok},y_o)(w_{ok}-z_\ell)^2\right\}\right]d\alpha(z_\ell)} \tag{4.3.8}$$

We make the same assumption as before,

$$\sum_{k_1=1}^{\infty} \frac{y_{jk_1}}{\lambda_{k_1}^2} < \infty \quad , \quad j=0,1,\ldots K. \tag{4.2.16}$$

Define

$$w_j = \lim_{k\to\infty} w_{jk} = \lim_{k\to\infty} \frac{(v_{jk},x)}{(v_{jk},y_j)} = \frac{(v_j,x)}{u_{jj}} \tag{4.3.9}$$

In any interval, Z_ℓ is just a constant. Hence, (4.2.16) guarantees that w_j exists with probability one. Then, taking the limit in (8) yields (by the bounded convergence theorem):

$$\lim_{k\to\infty} T_j(\underline{x}) = \frac{p_j}{p_o} \frac{e^{+\frac{1}{2}u_{jj}w_j^2}}{e^{+\frac{1}{2}u_{oo}w_o^2}} \frac{\int_{-\infty}^{+\infty} e^{-\frac{1}{2}u_{jj}(w_j-z_\ell)^2} d\alpha(z_\ell)}{\int_{-\infty}^{+\infty} e^{-\frac{1}{2}u_{oo}(w_o-z_\ell)^2} d\alpha(z_\ell)} \tag{4.3.10}$$

with probability one. The v_j's are again the (unique) solutions to the integral equation (4.2.18) and, since R(t) is assumed to be strictly positive definite, we have $u_{jj} > o$, j=0,...K.

For some of the cases we will discuss, it is more convenient to use the logarithm of the test function. Defining

$$L_j(w_j) = \sqrt{2\pi/u_{jj}} \int_{-\infty}^{+\infty} g(w_j - z_\ell; 1/\sqrt{u_{jj}}) d\alpha(z_\ell), \quad j=0,...K, \tag{4.3.11}$$

we have

$$\ln(T_j(w_j,w_o)) = \ln(p_j/p_o) + \frac{1}{2}(u_{jj}w_j^2 - u_{oo}w_o^2) + \ln L_j(w_j) - \ln L_o(w_o) . \tag{4.3.12}$$

The test procedure is to pass the received waveform $x(t), \ell \leq t \leq \ell + 1$, through K+1 matched filters. The gain of the j-th filter is $1/u_{jj}$, and the output (equation (9)) is w_j. The K+1 w_j are then used to evaluate the test function in (10) or (12), with the signal corresponding to the largest value being announced. This is the one-stage procedure to minimize the probability of error when $\alpha(z)$ is known. To estimate the test function in the nonsupervisory mode when $\alpha(z)$ is unknown, we first obtain the density function of the w_j.

Consider the output of the j-th matched filter in any interval. Assuming hypothesis i is active, we have

$$w_j = \frac{(v_j,x)}{u_{jj}} = \frac{1}{u_{jj}} \int_0^1 v_j(t)(n(t)+zy_i(t)dt . \tag{4.3.13}$$

w_j, given the value of z, is a gaussian random variable with mean value

$(zu_{ji})/u_{jj}$ and standard deviation $1/\sqrt{u_{jj}}$. The density function of w_j given that hypothesis i is in effect is

$$f_j(w_j|\ y=y_i) = \int_{-\infty}^{+\infty} g(w_j - z\,\frac{u_{ji}}{u_{jj}}\ ;\ 1/\sqrt{u_{jj}})d\alpha(z)\ , \qquad (4.3.14)$$

and averaging over all hypotheses we obtain

$$f_j(w_j) = \sum_{i=0}^{K} p_i \int_{-\infty}^{+\infty} g(w_j - z\,\frac{u_{ji}}{u_{jj}}\ ;\ 1/\sqrt{u_{jj}})\ d\alpha(z),\ j=0,\ldots K. \qquad (4.3.15)$$

We have dropped the ℓ subscript since the sequence $\{Z_\ell\}$ is assumed to be stationary.

4.3a Orthogonal Signals

We first consider the case of orthogonal signals. With $u_{ji}=0$ for $j\neq i$, (15) reduces to

$$f_j(w_j) = p_j \int_{-\infty}^{+\infty} g(w_j - z;\ 1/\sqrt{u_{jj}})d\alpha(z)$$

$$+ (1-p_j)g(w_j;\ 1/\sqrt{u_{jj}}). \qquad (4.3.16)$$

In view of the definition of $L_j(w_j)$, we can write

$$L_j(w_j) = \frac{\sqrt{2\pi/u_{jj}}}{p_j}\,(f_j(w_j) - (1-p_j)g(w_j;\ 1/\sqrt{u_{jj}}))\ . \qquad (4.3.17)$$

Since u_{jj} and p_j are known, the unknown part of $\ln(T_j)$ is $\ln(L_j(w_j))$. From (17), the problem of estimating $L_j(w_j)$ reduces to estimating $f_j(w_j)$

and then subtracting off the known quantities.[1] Clearly, we can do this under a nonsupervisory condition with either the L_2 series or the eigenfunction representation.

For the L_2 series approach, we expand each of the $f_j(w_j)$ in a series as in section 2.4. Then, assuming the r-th absolute moment of the random variable Z exists ($r \geq 3$), the sequence Z_ℓ is M-dependent, and that R(t) satisfies condition B, we can apply Corollary 2.4.1 to obtain

$$E_n\left\{\left[\hat{f}_{jn}(W_{jn})-f_j(W_{jn})\right]^2\right\} = O(1/n^{(r-2)/r}), \qquad (4.3.18)$$

where W_{jn} represents the output of the j-th filter during the n-th interval. This bound is then used to dominate the quantity $P_{en}-P_e$. Since (18) implies convergence in probability, one could proceed to dominate the expected value of $\ln^2 \hat{L}_{jn}(W_{jn})$ as in the previous section. In this manner, one could make a (probability) statement about the convergence of P_{en} to P_e in analogy to the case in sub-section 4.2b.

A block diagram of the procedure for estimating $L_j(w_j)$ is given in Figure 2. As indicated, the output of each matched filter is fed into q+1 devices to evaluate the first q+1 coefficients. The σ_i for each of the L_2 series represents an arbitrary constant. If we picked them all equal, we would not need q+1 coefficient generators for each

[1] The structure here is identical to the problem of binary on-off communication through a random channel. In both cases, we are testing a composite hypothesis vs. a specified alternative. The distribution of the composite hypothesis is estimated by estimating the overall distribution and then subtracting the known quantities.

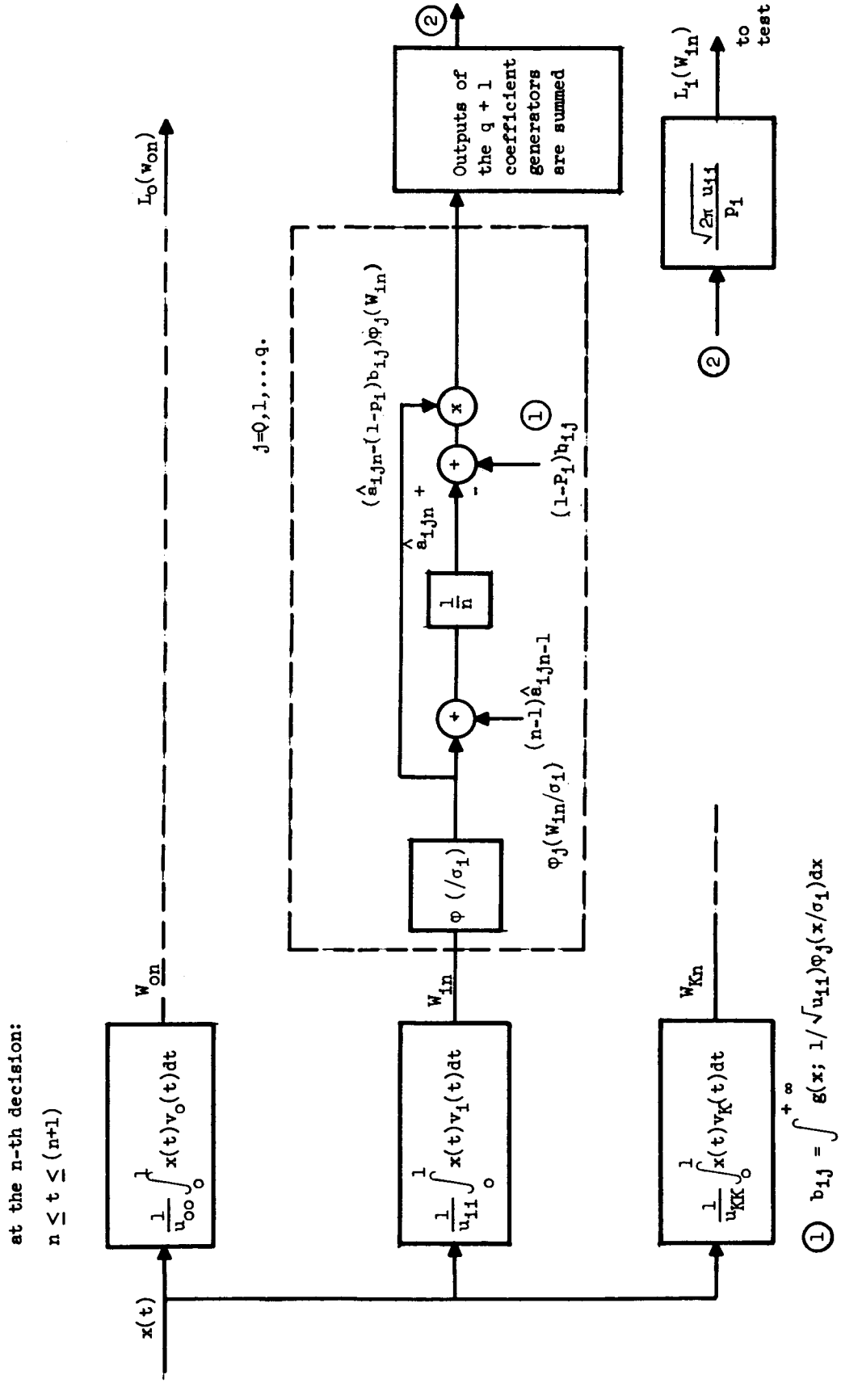

Figure 2. The L_2 series procedure for estimating $L_j(w_j)$—the case of orthogonal signals.

filter; the observations W_{jn}, j=0, 1...K, could be processed serially by the same q+1 devices. We can also eliminate storing the first q+1 Hermite functions by use of the recurrence formula

$$\varphi_{j+1}(w/\sigma) = \frac{2w}{\sigma}\varphi_j(w/\sigma) + 2j\,\varphi_{j-1}(w/\sigma) .$$

Since we have truncated the estimate of the series, there will be an asymptotic error as given by (2.4.37). We have,

$$\lim_{n\to\infty} E_n\left\{\left[(f_j(W_{jn}) - \tilde{f}_{jn}(W_{jn}))^2\right]\right\} \le c_2\sqrt{B_6}\left\{\frac{1}{(\frac{r}{2}-1)\,q^{(r/2)-1}} + \frac{1}{q^{r/2}}\right\} .$$

The constant c_2 is $c_1/(\pi^{1/4}\,\sigma_j^2)$. Recall that c_1 is Cramer's bound and σ_j is the arbitrary constant. B_6 is the L_2 norm of the function given below (2.4.21). Both the signal-to-noise ratio u_{jj} and σ_j enter into B_6.

We now use the eigenfunction representation for the same problem of orthogonal signals. From (16) and (2.5.11), $f_j(w_j)$ can be expressed as:

$$f_j(w_j) = \frac{1}{s_j(w_j)}\sum_{i=0}^{\infty}(p_j\xi_j^i\,d_{ji} + (1-p_j)b_{ji})\varphi_i(w_j/\gamma_j) \qquad (4.3.19)$$

where

$$s_j(w_j) = \frac{1}{\sqrt{\xi_j}}\; e^{-\frac{1}{2}\,\frac{w_j^2 u_{jj}}{\gamma_j^2\sigma_j^2}}$$

$$\xi_j^2 = \frac{u_{jj}\,\sigma_j^2-1}{u_{jj}\,\sigma_j^2+1} \tag{4.3.20}$$

$$\gamma_j^2 = \frac{(u_{jj}\sigma_j^2-1)(u_{jj}\sigma_j^2+1)}{u_{jj}^2\sigma_j^2}$$

and

$$d_{ji} = \int \varphi_i(z/\sigma_j)d\alpha(z) \ . \tag{4.3.21}$$

The b_{ji} are the Fourier coefficients of $s_j(w)g(w;1/\sqrt{u_{jj}})$,

$$b_{ji} = \int \varphi_i(w/\sigma_j)s_j(w)g(;1/\sqrt{u_{jj}})dw \ . \tag{4.3.22}$$

The constant σ_j^2 is chosen to be greater than $1/u_{jj}$.

The part of the test function which we want to estimate is $\ln L_j(w_j)$. From the definition of L_j we have the expansion

$$s_j(w_j)L_j(w_j) = \sqrt{2\pi/u_{jj}} \sum_{i=0}^{\infty} \xi_j^i \, d_{ji} \, \varphi_i(w_j/\gamma_j) \ , \tag{4.3.23}$$

and can write for (12)

$$\ln(T_j(w_j,w_o)) = \ln(p_j/p_o) + \frac{1}{2}(u_{jj}w_j^2 - u_{oo}\,w_o^2)$$

$$+ \ln(s_j(w_j)L\,(w_j)) - \ln(s_o(w_o)L_o(w_o))$$

$$- \ln s_j(w_j) + \ln s_o(w_o). \tag{4.3.24}$$

The error in the estimate of (24) is

$$\ln T_j - \ln\hat{T}_{jn} = \ln(s_jL_j) - \ln(s_j\hat{L}_{jn}) - \ln(s_oL_o) + \ln(s_o\hat{L}_{on}) \ , \tag{4.3.25}$$

which we can also write as

$$\ln T_j - \ln T_{jn} = \ln\left(1 + \frac{s_o \hat{L}_{on} - s_o L_o}{s_o L_o}\right) - \ln\left(1 + \frac{s_j \hat{L}_{jn} - s_j L_j}{s_j L_j}\right) \tag{4.3.26}$$

In analogy to section 2.5, we take the estimate of $s_j(w_j)L_j(w_j)$ as

$$s_j(w_j)\hat{L}_{jn}(w_j) = \sum_{i=0}^{q(n)} \xi_j^i \hat{d}_{jin}\, \varphi_i(w_j/\gamma_j) \tag{4.3.27}$$

where the estimates of the coefficients

$$\xi_j^i \hat{d}_{jin} = \frac{1}{p_j}\left\{\frac{1}{n}\sum_{\ell=1}^{n} \varphi_i(W_{j\ell}/\gamma_j)\, s(W_{j\ell}) - (1-p_j)b_{ji}\right\}.$$

differ from those in section 2.5 only by the $(1-p_j)b_{ji}$ term.

If the sequence of random variables $\{Z_\ell\}$ is M-dependent and if R(t) satisfies condition B, then, from Corollary 2.5.1, the sequence of estimates $s_j\hat{L}_{jn}$ converges in mean-square (uniformly in w_j) to s_jL_j. The rate is $\ln^2 n/n$ with the bound given by (2.5.34). Since we have convergence in probability, we could again make a probability statement concerning the convergence of P_{en} to P_e.

We remark that if the orthogonal signals $y_j(t)$ have equal energy then $u_{jj} = \ldots = u_{oo}$, and the $f_j(w_j)$ are essentially the same.[1] In this situation we need only use the output of one filter to estimate the

[1] The $f_j(w_j)$ would be identical if the signals were also equi-probable.

coefficients.

4.3b Bipolar Signals

Some interesting problems arise when we are testing between two hypotheses and the signals are bipolar.

The signals $y_0(t)$ and $y_1(t)$ are said to be bipolar (or antipodal) if $y_0(t) = -y_1(t)$. In this case, we need only one matched filter as $v_0(t) = -v_1(t)$. The density function of the observation (equation (15)) is

$$f(w_o) = p_o \int g(w_o - z;\ 1/\sqrt{u_{oo}})d\alpha(z) + p_1 \int g(w_o + z;\ 1/\sqrt{u_{oo}})d\alpha(z) \tag{4.3.28}$$

since $u_{oo}=u_{11}=-u_{o1}$. With $w_o = -w_1$, the test function, as given by (10) becomes

$$T(w_o) = \frac{p_1}{p_o} \frac{\int g(w_o - z;\ 1/\sqrt{u_{oo}})d\alpha(z)}{\int g(w_o + z;\ 1/\sqrt{u_{oo}})d\alpha(z)} . \tag{4.3.29}$$

Let

$$\tilde{L}_o(w_o) = s_o(w_o) \int g(w_o - z;\ 1/\sqrt{u_{oo}})d\alpha(z) \tag{4.3.30}$$

$$\tilde{L}_1(w_o) = s_o(w_o) \int g(w_o + z;\ 1/\sqrt{u_{oo}})d\alpha(z) ,$$

where $s_o(w_o)$ is defined as in (20). A procedure equivalent to comparing $T(w_o)$ to a threshold of one is to evaluate

$$\tilde{T}(w_o) = p_1 \tilde{L}_1(w_o) - p_o \tilde{L}_o(w_o) \tag{4.3.31}$$

and compare it to a zero threshold.

We need to assume that $d\alpha(z) \neq d\alpha(-z)$. If $d\alpha(z) = d\alpha(-z)$ then $\tilde{L}_1(w_o) = \tilde{L}_o(w_o)$ and there is no basis for a decision. Hence, if the density function exists ($d\alpha(z) = \alpha'(z)dz$), we assume that it is not an even function.

Multiplying (28) by $s_o(w_o)$ results in the expansion

$$s_o(w_o)f(w_o) = \sum \xi_o^j \, d_j(p_1+p_o(-1)^j) \, \varphi_j(w_o/\gamma_o) \tag{4.3.32}$$

while the test function can be written as

$$\tilde{T}(w_o) = \sum \xi_o^j d_j(p_1-p_o(-1)^j) \, \varphi_j(w_o/\gamma_o) \, . \tag{4.3.33}$$

The coefficients d_j are defined in (21). We have used the property that the Hermite functions are even or odd functions (depending on whether the index j is even or odd) to obtain (32) and (33).

We observe from (32) that if $p_1 \neq p_o$, there is no problem in estimating the product $(\xi^j d_j)$ which is then used to estimate $\tilde{T}(w_o)$. For bipolar signals, however, it is certainly reasonable to take $p_o = p_1 = \frac{1}{2}$. Then, (32) and (33) become

$$s_o(w_o)f(w_o) = \sum_{j=0}^{\infty} \xi_o^{2j} \, d_{2j} \, \varphi_{2j}(w_o/\gamma_o) \tag{4.3.34}$$

$$\tilde{T}(w_o) = \sum_{j=0}^{\infty} \xi_o^{2j+1} \, d_{2j+1} \, \varphi_{2j+1}(w_o/\gamma_o) \tag{4.3.35}$$

Hence, we can only estimate the even coefficients, while the test function depends on the odd coefficients.

There is one set of circumstances where this difficulty can be overcome.[1] Suppose the density function of z exists,

$$d\alpha(z) = \alpha'(z)dz \ ,$$

and that $\alpha'(z)$ is L_2. Further, assume that the random variable z can only take on positive values, $\alpha'(z) = 0$ for $z < 0$. Then, the even part of $\alpha'(z)$ uniquely determines the odd part:

$$\alpha'_e(z) = \frac{\alpha'(z)+\alpha'(-z)}{2} = \sum_{j=0}^{\infty} d_{2j}\, \varphi_j(z/\sigma_o) \qquad (4.3.36)$$

and

$$\alpha'_o = \frac{\alpha'(z)-\alpha'(-z)}{2} = \alpha'_e(z) \quad \text{for} \quad z > 0$$

$$= -\alpha'_e(z) \quad \text{for} \quad z < 0 \ .$$

From (30) and (31), the test function can be written as

$$\tilde{T}(w_o) = s(w_o) \int_{-\infty}^{+\infty} g(w_o - z;\ 1/\sqrt{u_{oo}}) \left[\frac{\alpha'(z)-\alpha'(-z)}{2}\right] dz$$

The procedure, then, would be to obtain estimates of the coefficients

---

[1] Operating in the supervisory mode, there is no difficulty since we can estimate all the coefficients, even and odd, by estimating the marginal density $f_o(w_o)$ or $f_1(-w_o)$. The even coefficients of f_o and f_1 are equal and the odd coefficients differ by a minus sign.

d_{2j} by estimating $(\xi^{2j} d_{2j})$ in (34), and then dividing by ξ^{2j}. The estimate of the test function would take the form

$$\tilde{T}_n(w_o) = s(w_o) \sum_{j=0}^{q(n)} \hat{d}_{2jn} \left\{ \int_o^{+\infty} g(w_o - z;\ 1/\sqrt{u_{oo}})\ \varphi_{2j}(z/\sigma_o) \right.$$

$$\left. - \int_{-\infty}^{0} g(w_o - z;\ 1/\sqrt{u_{oo}})\ \varphi_{2j}(z/\sigma_o) \right\} . \qquad (4.3.38)$$

The convergence of (38) to $\tilde{T}(w_o)$ is not given by the $\ln^2 n/n$ rate since we are, in effect, estimating $\alpha'(z)$ and not $f(x)$. As indicated above, using the eigenfunction representation, we can "pick off" the d_j's. The L_2 series can also be used to estimate $\alpha'(z)$. The estimation procedure, however, is more complicated. This will be discussed in section 4.4

4.3c Arbitrary Signals

The distinction between the supervisory and nonsupervisory modes is more marked in the case of general signals. Define the constants

$$\sigma_{ji} = \frac{\sigma_j\, u_{ji}}{u_{jj}}$$

$$\xi_{ji}^2 = \frac{u_{jj}\, \sigma_{ji}^2 - 1}{u_{jj}\, \sigma_{ji}^2 + 1} \qquad (4.3.39)$$

$$\gamma_{ji}^2 = \frac{(u_{jj}\sigma_{ji}^2 - 1)(u_{jj}\sigma_{ji}^2 + 1)}{u_{jj}^2\, \sigma_{ji}^2}$$

and the functions

$$s_{ji}(w_j) = \frac{1}{\sqrt{\xi_{ji}}} \exp\left\{ -\frac{1}{2} \frac{w_j^2 u_{jj}}{\gamma_{ji}^2 \sigma_{ji}^2} \right\} \qquad (4.3.40)$$

where the indices $i,j = 0,1 \ldots K$. The $(K+1)$ σ_j are chosen so that $\sigma_{ji}^2 > u_{jj}$. From (2.5.6), we have the expansion

$$s_{ji}(w_j) g(w_j - y;\ 1/\sqrt{u_{jj}}) = \sum_{k=0}^{\infty} \xi_{ji}^k \varphi_k(w_j/\gamma_{ji})\, \varphi_k(y/\sigma_{ji})$$

or,

$$s_{ji}(w_j) g(w_j - z\,\frac{u_{ji}}{u_{jj}};\ 1/\sqrt{u_{jj}}) = \sum_{k=0}^{\infty} \xi_{ji}^k \varphi_k(w_j/\gamma_{ji}) \varphi_k(z/\sigma_j) \qquad (4.3.41)$$

Defining the coefficients

$$d_{jk} = \int_{-\infty}^{+\infty} \varphi_k(z/\sigma_j)\, d\alpha(z) , \qquad (4.3.42)$$

the density function of the output of the j-th matched filter (equation (15)) becomes

$$f_j(w_j) = \sum_{i=0}^{K} p_i \frac{1}{s_{ji}(w_j)} \sum_{k=0}^{\infty} \xi_{ji}^k\, d_{jk}\, \varphi_k(w_j/\gamma_{ji}) . \qquad (4.3.43)$$

It is the d_{jk} which are needed to estimate the unknown part of the test function, $\ln L_j(w_j)$. Repeating the definition of $L_j(w_j)$,

$$L_j(w_j) = \sqrt{2\pi/u_{jj}} \int_{-\infty}^{+\infty} g(w_j - z; 1/\sqrt{u_{jj}})\, d\alpha(z) , \qquad (4.3.11)$$

we use (42) to obtain

$$L_j(w_j) = \frac{\sqrt{2\pi/u_{jj}}}{s_{jj}(w_j)} \sum_{k=0}^{\infty} d_{jk}\, \xi_{jj}^{k}\, \varphi_k(w_j/\gamma_{jj}) \ . \qquad (4.3.44)$$

The difficulty in estimating d_{jk} is that the unknowns in the overall density function no longer appear as the coefficients of an orthogonal expansion. Of course, in the supervisory mode, there is no need to consider the overall density function $f_j(w_j)$. By working with the individual densities in (43), we have the usual orthogonal expansion and can update the estimates of d_{jk} at each stage. For the nonsupervisory problem, we can use a procedure analogous to that given in section 2.6 to obtain approximate estimates for a finite number of the d_{jk}. By defining

$$\psi_{jk}(w_j) = \sum_{i=0}^{K} p_i \ \frac{1}{s_{ji}(w_j)} \ \xi_{ji}^{k} \varphi_k(w_j/\gamma_{ji}), \qquad (4.3.45)$$

$f_j(w_j)$ can be expressed as

$$f_j(w_j) = \sum_{k=0}^{\infty} d_{jk}\, \psi_{jk}(w_j) \ . \qquad (4.3.46)$$

It can be shown that the functions $\psi_{jk}(w_j)$, k=0,1,...J-1, are linearly independent. Then, we can construct (as in section 2.6) a set of functions $\psi_{jk}^{\perp}(w_j)$ which satisfy

$$\int_{-\infty}^{+\infty} \psi_{jk}^{\perp}(w)\psi_{jm}(w)dw = \begin{cases} 0, & m \neq k \\ 1, & m = k \end{cases} \qquad (4.3.47)$$

where k,m=0,1,...J-1. As an estimate of d_{jk}, k=0,1,...J-1, we can take

$$\hat{d}_{jkn} = \frac{1}{n} \sum_{\ell=1}^{n} \psi_{jk}^{\perp}(W_{j\ell}). \tag{4.3.48}$$

These estimates are biased

$$E(d_{jkn}) - d_{jk} = \sum_{i=J}^{\infty} d_{ji} \int_{-\infty}^{+\infty} \psi_{jk}^{\perp}(w)\, \psi_{ji}(w)dw \ . \tag{4.3.49}$$

The variance calculations are unwieldy and we will not discuss them.

The difficulties encountered in the case of arbitrary signals carry over when the received waveform is

$$x(t) = N(t) + Z(t)y_i(t);\ \ell \leq t \leq (\ell+1),\ i = 0,...K. \tag{4.3.2}$$

We define the time samples $z(t_{k_1})y_i(t_{k_1})$ as $z_{k_1}y_{ik_1}$, and let $\underline{u}_i$ be the vector of time samples. The density function of the k samples, given that the i-th hypothesis is active, is

$$f_i(\underline{x}) = \sum_{j_1 \cdots j_k} d_{j_1 \cdots j_k} \int \cdots \int g_k(\underline{x}-\underline{u}_i;A)\varphi_j(u_{i_1}/y_{i_1}\sigma_1)\cdots$$

$$\varphi_{j_k}(u_{i_k}/y_{i_k}\sigma_k)\ \frac{du_{i_1}\cdots du_{i_k}}{y_{i_1}\cdots y_{i_k}} \ . \tag{4.3.50}$$

For simplicity, we have assumed that the density function of $\alpha(\underline{z})$ exists and is L_2, and have defined the coefficients as

$$d_{j_1 \ldots j_k} = \cdots \int \cdots \int \varphi_{j_1}(z_1/\sigma_1) \cdots \varphi_{jk}(z_k/\sigma_k)\alpha'(z_1, \ldots z_k)dz_1 \ldots d_{z_k} \,. \tag{4.3.51}$$

The test procedure involves evaluating each of the $f_i(\underline{x})$ and then comparing them as in section 1.2.

If the noise samples in each interval are independent, A is diagonal and the problem is a direct extension of the one above. For the supervisory mode, each $s_i(\underline{x})f_i(\underline{x})$ can be expressed in an orthogonal series. In the nonsupervisory mode, (50) is summed over i and one can obtain the k-variate analog of (46).

When A is not diagonal, $f_i(\underline{x})$ can not be expressed in an orthogonal series with the $d_{j_1 \ldots j_k}$ defined as in (51). Be redefining the coefficients so as to depend on the index i (let $\sigma_1 \ldots \sigma_k$ depend on i), one could express each $f_i(\underline{x})$ in the same orthogonal series. Then, in the supervisory mode, the coefficients could be estimated as in section 3.3.c. For the nonsupervisory mode, this representation would only be useful for the case of bipolar signals.

4.4 UNBOUNDED LOSS FUNCTIONS FOR THE CASE OF BIPOLAR SIGNALS

For the case of equi-probable bipolar signals, we had to assume that the random variable Z could only take on positive values in order to consistently estimate the test function in the nonsupervisory mode. Under the same assumption on Z, we now give another formulation which is illustrative of a class of problems.

Equation (4.3.28) can be written as

$$f(w_o) = \int_{-\infty}^{+\infty} g(w_o - z; 1/\sqrt{u_{oo}}) \left(\frac{d\alpha(z) - d\alpha(-z)}{2}\right) . \qquad (4.4.1)$$

The output of the matched filter is $W_o = N_o \pm Z$, depending on whether $y_o(t)$ or $y_1(t) = -y_o(t)$ was transmitted. N_o is the (derived) gaussian noise sample with variance equal to $1/u_{oo}$. Deciding between y_o or $-y_o$ is equivalent to deciding whether the prefixing on the (non-negative) random variable Z is + or -. Rewriting (1) as

$$f(w_o) = \int_{-\infty}^{+\infty} g(w_o - z; 1/\sqrt{u_{oo}}) d\alpha_e(z) , \qquad (4.4.2)$$

we can think of Z as a random variable with the symmetric distribution $\alpha_e(z)$. Z will be positive (negative) only if $y_o(-y_o)$ is transmitted. Hence, the decision problem is equivalent to deciding whether Z is positive or negative.

If Z is close to zero, it is more difficult to distinguish between the two hypotheses. The penalty of an incorrect decision should, accordingly, be small. On the otherhand, if Z is large and we make an incorrect decision, the loss should be high. As one possible loss function, we take the loss equal to the magnitude of Z for an incorrect decision and equal to zero if we decide correctly.

Using the notation of sub-section 1.2d (with z and w_o in place of λ and x), the loss function $L(t(w_o), z)$ is:

$$\begin{aligned} L(0,z) &= 0 \quad \text{if} \quad z > 0 \\ &= -z \quad \text{if} \quad z < 0 \\ L(1,z) &= z \quad \text{if} \quad z > 0 \\ &= 0 \quad \text{if} \quad z < 0 \end{aligned} \tag{4.4.3}$$

We announce $y_o(t)$ if $t(w_o) = 0$ and $-y_o(t)$ if $t(w_o)=1$.

Defining

$$b(z) = L(0,z) - L(1,z) = -z \, , \tag{4.4.4}$$

from (1.2.49), the test function is

$$T_{\alpha_e}(w_o) = \int_{-\infty}^{+\infty} b(z) g(w_o - z; 1/\sqrt{u_{oo}}) d\alpha_e(z)$$

The procedure which minimizes the average risk is to set

$$t(w_o) = 1 \text{ if } T_{\alpha_e}(w_o) \geq 0$$

$$= 0 \text{ otherwise.}$$

Substituting for $b(z)$ yields:

$$\begin{aligned} T_{\alpha_e}(w_o) &= -\int_{-\infty}^{+\infty} z g(w_o - z; 1/\sqrt{u_{oo}}) d\alpha_e(z) \\ &= -\left(w_o f(w_o) + \frac{1}{u_{oo}} f'(w_o)\right) . \end{aligned} \tag{4.4.5}$$

Hence, the test function does not depend explicitly on $\alpha_e(z)$. However, we now require estimates of the derivative of $f(w_o)$. To estimate the derivative, the techniques given in Chapter 2 are applicable. By differentiating the estimates of $f(w_o)$, we can obtain consistent estimates

of $f'(w_o)$.[1] As would be expected, convergence of these estimates takes place at a slower rate than the estimates of the density function. Then, assuming $\int z^2 d\alpha_e(z) < \infty$, Corollary 1.2.4 can be used to bound the difference in risks.

If $b(z)$ is a k-th order polynomial in z, the test function will depend on $f(w_o)$ and its first k derivatives.[2] Even when the test function can not be written as a polynomial, the problem of finding consistent estimates of $T\alpha_e(w_o)$ is, in principal, easier than the situation encountered in the last section for the case of arbitrary signals—the point being that here, the density function under either hypothesis ($Z > 0$ or $Z < 0$) is the same and is given by (2). Consequently, we can estimate $\alpha_e(z)$ and form an estimate of the test function by taking

$$\hat{T\alpha}_e n(w_o) = \int_{-\infty}^{+\infty} b(z)\, g(w_o - z; 1/\sqrt{u_{oo}})\, d\hat{\alpha}_{en}(z) \quad . \tag{4.4.6}$$

We briefly discuss how to estimate the distribution or, for simplicity, the density $\alpha_e'(z)$. We assume $\alpha_e'(z)$ is L_2. Using the eigenfunction representation, we first form the estimate of $\xi^j d_j$ as in section 2.5, and then divide by the known ξ^j. The estimate of $\alpha_e'(z)$ is taken as

---

[1] For the method of 2.3, we have verified this only with the gaussian kernel. To use the eigenfunction representation, we estimate the product $s(w_o)f'(w_o)$.

[2] For a discussion of this point, and other densities (besides the gaussian) which have this property, see [37].

$$\hat{\alpha}'_{en}(z) = \sum_{j}^{q(n)} \hat{d}_{jn} \varphi_j(z/\sigma_1) \ .$$

Using the L_2 series for $f(x)$, we define the estimate implicity by

$$\begin{aligned} \hat{f}_n(w_o) &= \int_{-\infty}^{+\infty} g(w_o - z; 1/u_{oo}) \hat{\alpha}'_{en}(z) dz \ . \\ &= \sum_{j=0}^{q(n)} \hat{a}_{jn} \varphi_j(w_o/\sigma_1) \end{aligned} \tag{4.4.7}$$

Since the Hermite functions, aside from a constant, are their own Fourier transforms, letting $M_{f_n}(v)$ be the transform of $\hat{f}_n(x)$, we obtain

$$M_{f_n}(v) = \sigma_1 \sum_{j=0}^{q(n)} (-1)^{j/2} \hat{a}_{jn} \varphi_j(\sigma_1 v). \tag{4.4.8}$$

Letting $M\alpha_e n(v)$ be the transform of $\hat{\alpha}'_{en}$, we have

$$M\alpha_e n(v) = e^{+v^2/2u_{oo}} M_{f_n}(v) \ . \tag{4.4.9}$$

To form the estimate of $\alpha'_e(z)$, take the inverse finite Fourier transform of (9),

$$\hat{\alpha}_{en}(z) = \int_{-B(n)}^{+B(n)} e^{+v^2/2u_{oo} - ivz} M_{fn}(v) \, dv \ ,$$

with $M_{f_n}(v)$ given by (8).

Both of these representations for $\alpha'_e(z)$ lead to consistent estimates and, under appropriate conditions on $b(z)$, the sequence of test functions defined in (6) can be shown to be consistent.

CHAPTER 5

SUMMARY

We have studied a particular empirical procedure and applied it to some problems in communication theory. The procedure utilizes all past observations to form an estimate of a test function which is then evaluated using only the present observation. This procedure is neither optimum nor asymptotically optimum when the sequence of observations is assumed to be dependent. Whether or not the sequence of observations is dependent, we have shown that if the sequence of test functions converges in mean-square to the one-stage test function, the difference in the risks (for the case of a bounded loss function) is dominated by a quantity proportional to the mean-square error in the estimate of the test function. These calculations (section 1.2), although straightforward, appear to be new.

In estimating the density function $f(x)$ from the sequence of dependent observations, we are able to dominate the mean-square error and hence, specify the rate of convergence of the estimate. The key relationship is the Mehler formula (2.2.12), or more generally, the Barrett-Lampard expansion (2.7.1). It appears that this particular application of the expansion has not been used before.

We have presented three methods of estimating the density $f(x)$. Varying amounts of information are required to apply and specify a rate of convergence for each of the techniques. A summary of assumptions

needed and rates of convergence are given in section 2.7 and at the end of 2.5. Of the three methods we have presented, the most interesting is the eigenfunction representation. To the best of our knowledge, this approach of estimating a density function obtained from a convolution and the particular solution of the eigenfunction problem in section 2.5 have not appeared in the literature.

The communication problems we have considered are those in which the unknowns enter linearly into the overall density function. In section 4.2, the unknowns consist of a finite set of parameters. In the other problems, an arbitrary distribution is taken as unknown and expressed in terms of a countable set of linear unknown parameters and known functions by using either of the two series methods.

The series methods we have studied may also be useful for other purposes. We have mentioned two; estimating a k-variate density function with the L_2 series (2.7 and 3.3b) and using the eigenfunction representation for the detection of random signals in gaussian noise with the number of terms to be used in the test function determined by a sequential procedure (4.1b).

APPENDIX A

THE HERMITE POLYNOMIALS

In this appendix, we review the definitions and properties of the Hermite polynomials and develop the relationships which we will need.

Common practice has been to use the notation $He_n(x)$ for the polynomials associated with the weight function $e^{-x^2/2}$, and $H_n(x)$ for the polynomials associated with the weight e^{-x^2}. This is the notation which we will adopt. The polynomials are related by (Erdelyi, [12] p. 268),

$$He_n(x) = 2^{-\frac{n}{2}} H_n(x/\sqrt{2}) \tag{A.1}$$

$$H_n(x) = 2^{n/2} He_n(\sqrt{2}x). \tag{A.2}$$

A.1 THE POLYNOMIALS $He_n(x)$

From Cramèr, [10], page 133, we define

$$He_n(x) = (-1)^n e^{x^2/2} \left(\frac{d}{dx}\right)^n e^{-x^2/2} \tag{A.3}$$

The first five polynomials are:

$$He_0(x) = 1 \qquad He_3(x) = x^3 - 3x \tag{A.4}$$

$$He_1(x) = x \qquad He_4(x) = x^4 - 6x^2 + 3$$

$$He_2(x) = x^2 - 1$$

The orthogonality relation is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} \, He_n(x) \, He_m(x) dx = n! \quad , \quad m = n$$
$$= 0 \quad , \quad m \neq n \tag{A.5}$$

One generating function is

$$e^{-t^2/2 + tx} = \sum_{0}^{\infty} \frac{t^v}{v!} \, He_v(x) \tag{A.6}$$

and another particularly useful expansion is given by

$$\frac{1}{\sqrt{1-\rho^2}} \, e^{-\left\{\frac{\rho^2 x^2 + \rho^2 y^2 - 2\rho xy}{2(1-\rho^2)}\right\}} = \sum_{0}^{\infty} \frac{\rho^v}{v!} \, He_v(x) \, He_v(y), \tag{A.7}$$

with (A.7) holding for $|\rho| < 1$.

We define (see Table of Symbols)

$$g(x-m;\sigma) = g\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\,\sigma} \, e^{-\frac{(x-m)^2}{2\sigma^2}} . \tag{A.8}$$

We have from (A.3)

$$\frac{d^n}{dx^n} \, g(x-m;\sigma) = \frac{(-1)^n}{\sigma^n} \, g\left(\frac{x-m}{\sigma}\right) He_n\left(\frac{x-m}{\sigma}\right) , \tag{A.9}$$

and from (A.5)

$$\int_{-\infty}^{+\infty} He_n\left(\frac{x-m}{\sigma}\right) He_\ell\left(\frac{x-m}{\sigma}\right) g\left(\frac{x-m}{\sigma}\right) dx = n! \quad , \ell = n$$
$$= 0 \quad , \ell \neq n . \tag{A.10}$$

A.2 THE POLYNOMIALS $H_n(x)$

From Erdelyi, [12], section 10.13, define:

$$H_n(x) = (-1)^n e^{x^2} \left(\frac{d}{dx}\right)^n e^{-x^2} \tag{A.11}$$

$H_n(x)$ is also given by

$$H_n(x) = n! \sum_{m=0}^{[n/2]} \frac{(-1)^m (2x)^{n-2m}}{m!\,(n-2m)!} \quad , \tag{A.12}$$

where $[n/2]$ denotes the largest integer $n/2$ or $(n-1)/2$. The corresponding five polynomials are:

$$H_0(x) = 1 \qquad H_3(x) = 8x^3 - 12x$$

$$H_1(x) = 2x \qquad H_4(x) = 16x^4 - 48x^2 + 12 \; .$$

$$H_2(x) = 4x^2 - 2 \tag{A.13}$$

The orthogonality relation is

$$\int_{-\infty}^{+\infty} H_n(x)\, H_m(x)\, e^{-x^2} dx = 2^n n! \sqrt{\pi}\; \delta_{mn} \tag{A.14}$$

and the corresponding generating functions for these polynomials are:

$$e^{-t^2+2tx} = \sum_{n=0}^{\infty} \frac{t^n}{n!} H_n(x) \tag{A.15}$$

$$\frac{1}{\sqrt{1-\rho^2}}\, e^{-\frac{\{x^2\rho^2+y^2\rho^2-2\rho xy\}}{1-\rho^2}} = \sum_{n=0}^{\infty} \frac{(\rho/2)^n}{n!} H_n(x)\, H_n(y), \; |\rho| < 1 \tag{A.16}$$

(A.16) is called Mehler's formula.

For the weight function

$$g^2(x/\sigma) = \frac{1}{(2\pi)\sigma^2}\, e^{-\frac{x^2}{\sigma^2}} \quad ,$$

we have from (A.11)

$$\left(\frac{d}{dx}\right)^n g^2(x/\sigma) = \frac{(-1)^n}{\sigma^n} H_n\left(\frac{x}{\sigma}\right) g^2(x/\sigma) \qquad (A.17)$$

and from (A.14)

$$\int_{-\infty}^{+\infty} H_n\left(\frac{x}{\sigma}\right) H_m\left(\frac{x}{\sigma}\right) e^{-x^2/\sigma^2} dx = \sigma\, 2^n n!\sqrt{\pi}\, \delta_{mn}. \qquad (A.18)$$

In terms of the $g(x/\sigma)$ notation, the orthogonality relation becomes

$$\int_{-\infty}^{+\infty} H_n(x/\sigma) H_m(x/\sigma) g^2(x/\sigma) dx = \frac{2^n n!}{\sqrt{4\pi\sigma^2}}\, \delta_{mn} \qquad (A.19)$$

The sequence of functions given by $g(x/\sigma)H_j(x/\sigma)$ is known to be a complete orthogonal system [38].

A.3 THE EXPANSION OF THE BIVARIATE GAUSSIAN DENSITY FUNCTION AND A LEMMA

An expansion for the bivariate gaussian density function is easily obtained by multiplying (A.7) through by $\frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(x^2+y^2)}{2}}$ and then substituting $x = \frac{x-m_1}{\sigma_1}$, $y = \frac{y-m_2}{\sigma_2}$. In terms of our $g(x)$ notation, the result is

$$\frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp - \frac{1}{2}\left\{\frac{\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}}{1-\rho^2}\right\}$$

$$= g\left(\frac{x-m_1}{\sigma_1}\right) g\left(\frac{y-m_2}{\sigma_2}\right) \sum_{n=0}^{\infty} \frac{\rho^n}{n!} He_n\left(\frac{x-m_1}{\sigma_1}\right) He_n\left(\frac{y-m_2}{\sigma_2}\right), \quad |\rho| < 1 . \qquad (A.20)$$

This result can also be derived from the bivariate gaussian characteristic

function by expanding the cross-product term in a power series. The double integral in the inversion formula then becomes two single integrals.[1]

We denote the bivariate guassian density function by $g_2(x-m_1,y-m_2;\sigma_1,\sigma_2,\rho)$. If the standard deviation of both variables is the same, we designate the density by $g_2(x-m_1,y-m_2;\sigma,\rho)$.

When integrating the bivariate density as given by (A.20), we will want to interchange the double integration and summation. This is easily justified by

Lemma A.1: With $|\rho| < 1$, it follows that

$$\int_{-\infty}^{y_1}\int_{-\infty}^{y_2} g_2(x_1,x_2,;\sigma_1,\sigma_2,\rho)dx_1dx_2 \quad =$$

$$= \sum_{n=0}^{\infty} \frac{\rho^n}{n!} \int_{-\infty}^{y_1} g(x_1/\sigma_1)He_n(x_1/\sigma_1)dx_1 \int_{-\infty}^{y_2} g(x_2/\sigma_2)He_n(x_2/\sigma_2)dx_2 \qquad (A.21)$$

Proof: Let

$$G(x_1,x_2) = \sum_{n=0}^{\infty} \frac{|\rho|^n}{n!} |He_n(x_1/\rho_1)||He_n(x_2/\sigma_2)|g(x_1/\sigma_1)g(x_2/\sigma_2) \qquad (A.22)$$

Then by the Lebesgue monotone convergence theorem,

$$\int_{-\infty}^{y_1}\int_{-\infty}^{y_2} G(x_1,x_2)dx_1dx_2 \quad =$$

---

[1] Stratonovich, R. L., Topics in Theory of Random Noise, Gordon and Breach, N. Y., 1963. Translated from the Russian by R. A. Silverman pp. 41-42.

$$= \sum_{n=0}^{\infty} \frac{|\rho|^n}{n!} \int_{-\infty}^{y_1} |He_n(x_1/\sigma_1)| g(x_1/\sigma_1) dx_1 \int_{-\infty}^{y_2} |He_n(x_2/\sigma_2)| g(x_2/\sigma_2) dx_2 \qquad (A.23)$$

Using (A.10) and the Schwarz inequality, we obtain

$$\int_{-\infty}^{y_1} |He_n(x_1/\sigma_1)| g(x_1/\sigma_1) dx_1 \leq \left\{ \int_{-\infty}^{y_1} He_n^2(x_1/\sigma_1) g(x_1/\sigma_1) dx_1 \int_{-\infty}^{y_1} g(x_1/\sigma_1) dx_1 \right\}^{\frac{1}{2}}$$

$$\leq \sqrt{n!} \qquad (A.24)$$

Since $|\rho| < 1$, (A.23) is dominated by

$$\int_{-\infty}^{y_1} \int_{-\infty}^{y_2} G(x_1,x_2) dx_1 dx_2 \leq \sum_{n=0}^{\infty} |\rho|^n = \frac{1}{1-|\rho|} \qquad (A.25)$$

Now define $g_N(x_1,x_2) = \sum_{n=0}^{N} \frac{\rho^n}{n!} g(x_1/\sigma_1) g(x_2/\sigma_2) He_n(x_1/\sigma_1) He_n(x_2/\sigma_2)$. Clearly, $g_N(x_1,x_2) \to g(x_1,x_2;\sigma,\rho)$ pointwise. For all N we have $g_N(x_1,x_2) \leq G(x_1,x_2)$ which we just showed was integrable. (A.21) then follows from the Lebesgue dominated convergence theorem.

A.4 MISCELLANEOUS RELATIONSHIPS

From Erdelyi, [12], p. 193, we have the fact that $H_n(x)$ is either an even or odd function, depending on the index being even or odd,

$$H_n(x) = (-1)^n H_n(-x) . \qquad (A.26)$$

From the same page we have:

$$H_{n+1}(x) - 2xH_n(x) + 2n H_{n-1}(x) = 0 \qquad (A.27)$$

$$\frac{d}{dx} H_n(x) = H_n'(x) = 2n H_{n-1}(x) . \qquad (A.28)$$

A uniform bound (in x) for the Hermite functions was given by Cramer. From Erdelyi, [12], p. 208, or Sansone, [38], p. 324, Cramer's bound is

$$e^{-x^2/2} \, |H_n(x)| \; < \; c_1 \sqrt{2^n n!} \; , \tag{A.29}$$

where the constant $c_1 = 1.086435$. Using (A.1) a bound in terms of the $He_n(x)$ polynomials is

$$e^{-x^2/2} \, |He_n(x)| \leq e^{-x^2/4} |He_n(x)| < c_1 \sqrt{n!} \tag{A.30}$$

The bound has been improved with c_1 replaced by $2^{1/4}/\sqrt{\pi}$.[1]

---

[1] Reuter, G. E. H., "On the Boundedness of the Hermite Orthogonal System," Journal of the London Math. Society, vol. 24, April 1949, pp. 159-160.

APPENDIX B

EVALUATION OF SOME INTEGRALS

In this appendix we calculate a number of integrals involving the Hermite polynomials. For our purposes, the most useful result is given in equation (B.10). Equations (B.15) and B.17) are of interest and have been included for the sake of completeness.

We shall evaluate the integrals starting from an integral appearing in Tables of Integral Transforms [13]. By way of verification, we also indicate different (and sometimes more direct) methods of obtaining the results.

From Erdelyi, et. al.,[13], page 290, number 17:

$$\int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(x-y)^2\right] He_n(\alpha x)dx = (2\pi)^{1/2}(1-\alpha^2)^{n/2} He_n\left[\frac{\alpha y}{(1-\alpha^2)^{1/2}}\right] . \tag{B.1}$$

It is easy to verify this integral for n=0,1. Then, integrate (B.1) by parts and use the relations

$$\frac{d}{dx} He_n(\alpha x) = n\alpha He_{n-1}(\alpha x)$$

$$\alpha x\, He_n(\alpha x) = He_{n+1}(\alpha x) + n\, He_{n-1}(\alpha x)$$

(which are derived from (A.27) and (A.28)) to express (B.1) in terms of integrals involving polynomials of order n-1 and n-2. (B.1) is then

verified by induction. The simple substitution $x=\bar{x}/\sigma$, $y=\bar{y}/\sigma$ gives

$$\int_{-\infty}^{+\infty} g(\frac{\bar{x}-\bar{y}}{\sigma})\ He_n(\frac{\alpha\bar{x}}{\sigma})d\bar{x} \;=\; (1-\alpha^2)^{\frac{n}{2}}\ He_n\left[\frac{\alpha\bar{y}}{\sigma(1-\alpha^2)^{1/2}}\right] \quad , \tag{B.2}$$

where again $g(\frac{\bar{x}-\bar{y}}{\sigma})$ denotes the gaussian density with mean $\bar{y}$ and standard deviation σ. Using the relationship between the two types of Hermite polynomials, (A.1), the integral is expressed in terms of H_n.

$$\int_{-\infty}^{+\infty} g(\frac{\bar{x}-\bar{y}}{\sigma})\ H_n\ (\frac{\alpha\bar{x}}{\sqrt{2}\sigma})d\bar{x} \;=\; (1-\alpha^2)^{n/2}\ \ H_n\left[\frac{\alpha\bar{y}}{\sqrt{2}\sigma(1-\alpha^2)^{1/2}}\right] \quad . \tag{B.3}$$

With $x=\bar{x}-m$, this becomes

$$\int_{-\infty}^{+\infty} g(\frac{x-y-m}{\sigma})\ H_n\left[\frac{\alpha(x-m)}{\sqrt{2}\ \sigma}\right] dx = (1-\alpha^2)^{n/2}\ H_n\left[\frac{\alpha y}{\sqrt{2}\sigma(1-\alpha^2)^{1/2}}\right] \quad . \tag{B.4}$$

We now want to evaluate the following integral.

$$I_1 \;=\; \int_{-\infty}^{+\infty} H_n(\frac{x-m_1}{\sigma_1})\ g(\frac{x-m_1}{\sigma_1})\ g(\frac{x-m_2}{\sigma_2})dx \quad . \tag{B.5}$$

Use the relation

$$g(\frac{x-m_1}{\sigma_1})\ g(\frac{x-m_2}{\sigma_2}) = g(\frac{m_1-m_2}{\sqrt{\sigma_1^2+\sigma_2^2}})\ g(\frac{x-ab^2}{b}) \quad , \tag{B.6}$$

where

$$a \;=\; \frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}$$

$$b^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

and I_1 becomes

$$I_1 = g\left(\frac{m_1-m_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) \int_{-\infty}^{+\infty} H_n\left(\frac{x-m_1}{\sigma_1}\right)\, g\left(\frac{x-ab^2}{b}\right)dx \ . \tag{B.7}$$

Make the following identifications

$$y = ab^2-m_1$$

$$\sigma = b$$

$$\alpha = \frac{\sqrt{2}b}{\sigma_1}$$

and substitute into (B.7).

$$I_1 = g\left(\frac{m_1-m_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) \int_{-\infty}^{+\infty} H_n\left[\frac{\alpha(x-m_1)}{\sqrt{2}\,\sigma}\right]\, g\left(\frac{x-y-m_1}{b}\right)dx \ . \tag{B.8}$$

This integral is given by (B.4)

$$I_1 = g\left(\frac{m_1-m_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) (1-\alpha^2)^{n/2}\ H_n\left[\frac{\alpha y}{\sqrt{2}\,\sigma(1-\alpha^2)^{1/2}}\right] \ . \tag{B.9}$$

Substitute for y,σ,α, and then for a and b. The result is

$$\begin{aligned} I_1 &= \int H_n\left(\frac{x-m_1}{\sigma_1}\right)\, g\left(\frac{x-m_1}{\sigma_1}\right)\, g\left(\frac{x-m_2}{\sigma_2}\right)dx \\ &= g\left(\frac{m_2-m_1}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) \left\{\frac{\sigma_1^2-\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right\}^{\frac{n}{2}}\ H_n\left[\frac{\sigma_1(m_2-m_1)}{\sqrt{\sigma_1^2+\sigma_2^2}\,\sqrt{\sigma_1^2-\sigma_2^2}}\right] \ . \end{aligned} \tag{B.10}$$

This integral exhibits a type of reproducing property which we will find useful in this study. The property which we refer to is the fact that the (gaussian) average of a Hermite polynomial of order n and its associated gaussian weight gives a Hermite polynomial of the same order

and a gaussian weight, with the resulting two functions having different arguments.

For the case where $\sigma_1=\sigma_2$, by using (A.12), (B.10) reduces to

$$I_1 = g\left(\frac{m_2-m_1}{\sqrt{2}\,\sigma_1}\right)\left(\frac{m_2-m_1}{\sigma_1}\right)^n . \tag{B.11}$$

As an alternate derivation, equations (B.10) and (B.11) can be obtained by using a form of the Weirstrass transform (Bilodeau [4]). Briefly, this method involves defining the transform of $\psi(y)$ by

$$\gamma(x) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-(x-y)^2} \psi(y)dy \tag{B.12}$$

and noting that (under suitable conditions on $\psi(y)$)

$$\frac{d^n}{dx^n}\gamma(x)\Big|_{x=0} = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} H_n(y)e^{-y^2} \psi(y)dy . \tag{B.13}$$

The evaluation of (B.12) is carried out by completing the squares of the product of appropriate guassian weights.

A generalization of (B.10) is the evaluation of

$$\int_{-\infty}^{+\infty} H_n\left(\frac{x-m_1}{\sigma_1}\right) g\left(\frac{x-m_2}{\sigma_2}\right) g\left(\frac{x-m_3}{\sigma_3}\right)dx . \tag{B.14}$$

We use (B.6) to manipulate (B.14) into a form so as to use (B.10). The result of this straightforward procedure is:

$$\int_{-\infty}^{+\infty} H_n\left(\frac{x-m_1}{\sigma_1}\right) g\left(\frac{x-m_2}{\sigma_2}\right) g\left(\frac{x-m_3}{\sigma_3}\right)dx$$

$$= g\left(\frac{m_2-m_3}{\sqrt{\sigma_2^2+\sigma_3^2}}\right) \left\{\frac{\sigma_2^2(\sigma_1^2-\sigma_3^2)+\sigma_3^2(\sigma_1^2-\sigma_2^2)}{\sigma_1^2(\sigma_2^2+\sigma_3^2)}\right\}^{\frac{n}{2}}$$

$$H_n\left[\frac{\sigma_3^2(m_2-m_1) + \sigma_2^2(m_3-m_1)}{\sqrt{\sigma_2^2+\sigma_3^2}\sqrt{\sigma_2^2(\sigma_1^2-\sigma_3^2)+\sigma_3^2(\sigma_1^2-\sigma_2^2)}}\right] . \qquad \text{(B.15)}$$

The polynomials $He_n(x)$ also exhibit the reproducing property analogous to (B.10). For the integral given below, set $\sigma_1 = \sqrt{2}\,\sigma_2$ and use (A.1):

$$\int_{-\infty}^{+\infty} He_n\left(\frac{y-x_2}{\sigma_2}\right) g\left(\frac{y-x_2}{\sigma_2}\right) g\left(\frac{y-x_3}{\sigma_3}\right)dy$$

$$= 2^{-\frac{n}{2}} \int_{-\infty}^{+\infty} H_n\left(\frac{y-x_2}{\sigma_1}\right) g\left(\frac{y-x_2}{\sigma_2}\right) g\left(\frac{y-x_3}{\sigma_3}\right) . \qquad \text{(B.16)}$$

This integral is given by (B.15). Upon substituting back for σ_1 and He we obtain the desired result:

$$\int He_n\left(\frac{y-x_2}{\sigma_2}\right) g\left(\frac{y-x_2}{\sigma_2}\right) g\left(\frac{y-x_3}{\sigma_3}\right)dy$$

$$= \left\{\frac{\sigma_2^2}{\sigma_2^2+\sigma_3^2}\right\}^{\frac{n}{2}} g\left(\frac{x_3-x_2}{\sqrt{\sigma_2^2+\sigma_3^2}}\right) He_n\left[\frac{x_3-x_2}{\sqrt{\sigma_2^2+\sigma_3^2}}\right] . \qquad \text{(B.17)}$$

Here, the argument of the resulting gaussian weight and Hermite polynomial are the same (cf. (B.10)).

This result can also easily be obtained from the integral

$$\int_{-\infty}^{+\infty} g\left(\frac{y-x_2}{\sigma_2}\right) g\left(\frac{y-x_3}{\sigma_3}\right)dy = g\left(\frac{x_3-x_2}{\sqrt{\sigma_2^2+\sigma_3^2}}\right)$$

and the following relationships:

$$\frac{d^n}{dx_3^n}\, g\left(\frac{x_3-x_2}{\sqrt{\sigma_2^2+\sigma_3^2}}\right) = \frac{(-1)^n}{(\sigma_2^2+\sigma_3^2)^{\frac{n}{2}}}\, g\left(\frac{x_3-x_2}{\sqrt{\sigma_2^2+\sigma_3^2}}\right) He_n\left(\frac{x_3-x_2}{\sqrt{\sigma_2^2+\sigma_3^2}}\right)$$

$$\frac{d^n}{dx_3^n}\, g\left(\frac{y-x_3}{\sigma_3}\right) = \frac{1}{\sigma_3^n}\, g\left(\frac{y-x_3}{\sigma_3}\right)\, He_n\left(\frac{y-x_3}{\sigma_3}\right)$$

$$He_n(-y) = (-1)^n\, He_n(y) \ .$$

Another integral which we will need is

$$I_2 = \int\int g(x_1-y_1;\sigma_1)g(x_2-y_2;\sigma_1)\ g_2(y_1-z_1,y_2-z_2;\sigma_2,\rho)dy_1dy_2 \qquad (B.18)$$

where $g_2(y_1-z_1,y_2-z_2;\rho)$ is the bivariate gaussian density with standard deviation σ_2 and correlation coefficient ρ. This integral can be evaluated by using (B.17) and the expansion (A.20). We note, however, that (B.18) represents the density function of the sum of two independent gaussian vectors. The resulting probability density function, which of course is gaussian, has a covariance matrix equal to

$$\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} + \begin{bmatrix} \sigma_2^2 & \rho\sigma_2^2 \\ \rho\sigma_2^2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2+\sigma_2^2 & \rho\sigma_2^2 \\ \rho\sigma_2^2 & \sigma_1^2+\sigma_2^2 \end{bmatrix}$$

Hence, defining

$$\gamma^2 = \sigma_1^2 + \sigma_2^2 \qquad (B.19)$$

$$\bar{\rho} = \rho\,\frac{\sigma_2^2}{\sigma_2^2+\sigma_1^2} \ ,$$

(B.18) is given by

$$I_2 = g_2(x_1-z_1,x_2-z_2;\gamma,\ \bar{\rho}) \ . \qquad (B.20)$$

APPENDIX C

THE GAUSSIAN KERNEL

By specializing the kernel in section 2.3 to $K(y)=g(y;1)$, we are able to perform some of the required integrations. This permits, for example, an exact expression for the bias and second moment of the estimate. It also leads to sharper bounds on the variance expression.

C.1 THE UNIVARIATE CASE

From section 2.3, the estimate of the density function is

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{\ell=1}^{n} K\left(\frac{x-X_\ell}{h}\right).$$

Take $K(x) = g(x;1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. The estimate is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{\ell=1}^{n} g(x-X_\ell;h). \qquad (C.1)$$

The mean value is given by

$$\begin{aligned} E\hat{f}_n(x) &= Eg(x-X;h) \\ &= \int g(x-y;h) \int g(y-z;\sigma)\, d\alpha(z)\, dy \\ &= \int g(x-z;\sqrt{\sigma^2+h^2})\, d\alpha(z). \qquad (C.2) \end{aligned}$$

The bias in the estimate is then

$$E\left\{\hat{f}_n(x) - f(x)\right\} = \int \left\{g(x-z;\sqrt{\sigma^2+h^2}) - g(x-z;\sigma)\right\} d\alpha(z). \qquad (C.3)$$

In (2.3.20), we defined the quantity

$$Q_{m-\ell} = \int\int K\left(\frac{x-y_1}{h}\right)K\left(\frac{x-y_2}{h}\right)$$

$$\left\{ f_{m-\ell}(y_1,y_2) - f(y_1)f(y_2) \right\} dy_1dy_2.$$

This expression, the third part of the variance bound, led to the $1/nh^2$ factor. Using the gaussian kernel we will be able to eliminate the $1/h^2$ term. Substuting for the kernel we obtain

$$\frac{Q_{m-\ell}}{h^2} = \int\int g(x-y_1;h)\,g(x-y_2;h) \left[f_{m-\ell}(y_1,y_2) - f(y_1)f(y_2) \right] dy_1dy_2 \tag{C.4}$$

Designate the first part of this expression by $Q_{m-\ell,1}$. Write out $f_{m-\ell}(y_1,y_2)$ and interchange the y and z integrations.

$$Q_{m-\ell,1} = \int\int d\alpha(z_1)\,d\alpha(z_2) \int\int g_2(y_1-z_1,y_2-z_2;\sigma,\rho_{m-\ell})$$

$$g(x-y_1;h)\,g(x-y_2;h)\,dy_1dy_2. \tag{C.5}$$

The double integration has been evaluated in the previous appendix. Define

$$\beta = \sqrt{\sigma^2+h^2}$$

$$\gamma_{m-\ell} = \rho_{m-\ell}\,\frac{\sigma^2}{\sigma^2+h^2}\,. \tag{C.6}$$

Then, from (B.20),

$$Q_{m-\ell,1} = \int\int d\alpha(z_1)\, d\alpha(z_2)\, g_2(x-z_1, x-z_2; \beta, \gamma_{m-\ell}). \qquad (C.7)$$

The second part of (C.4) is just the square of (C.2). Combining (C.7) and (C.2) we have,

$$\frac{Q_{m-\ell}}{h^2} = \int_{z_1}\int_{z_2} d\alpha(z_1)\, d\alpha(z_2) \Big\{ g_2(x-z_1, x-z_2; \beta, \gamma_{m-\ell}) - g(x-z_1;\beta)\, g(x-z_2;\beta) \Big\}. \qquad (C.8)$$

Noting that $|\gamma_{m-\ell}| \leq |\rho_{m-\ell}| < 1$ for $m \neq \ell$, we use the Mehler formula and Cramer's bound to obtain

$$|Q_{m-\ell}| \leq h^2 \frac{|\gamma_{m-\ell}|}{1-|\gamma_{m-\ell}|} \frac{c_1^2}{2\pi\beta} \leq \frac{h^2|\rho_{m-\ell}|}{1-|\rho_{m-\ell}|} \frac{c_1^2}{2\pi\sigma}, \qquad (C.9)$$

which is the result quoted in section 2.3, equation (2.3.24).

Note that we have taken the $\{Z_\ell\}$ as independent. The extension to M-dependent variables is straightfoward. An additional term is added to $Q_{m-\ell}/h^2$ which, after performing the y integrations, is given by

$$\int\int g_2(x-z_1, x-z_2; \beta, \gamma_{m-\ell}) \left[d\alpha_{m-\ell}(z_1, z_2) - d\alpha(z_1)\, d\alpha(z_2) \right]. \qquad (C.10)$$

This contribution to the $V(f_n(x))$ expression can be bounded by

$$\frac{4}{n} \frac{(M-1)}{2\pi\sigma^2(1-\rho_*)}.$$

The result in (C.5) enables us to get an exact expression for the second moment of the estimate. Using this and essentially (C.2), we have

$$E\{\hat{f}_n(x)\}^2 =$$

$$E\left\{g^2(x-X;h)\right\} + \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n} E\left\{g(x-X_\ell;h)\,g(x-X_m;h)\right\}$$

$$= \frac{1}{nh2\sqrt{\pi}}\int g(x-z;\sqrt{\sigma^2+h^2/2})\,d\alpha(z)$$

$$+ \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}\int\int d\alpha(z_1)\,d\alpha(z_2)\,g_2(x-z_1,x-z_2;\beta,\gamma_{m-\ell}). \qquad (C.11)$$

C.2 MEAN INTEGRATED SQUARE ERROR

In section 2.3d, we stated that the MISE was $= O(1/n^{4/5})$. We now obtain this bound assuming the $\{Z_\ell\}$ are independent and that the autocorrelation function satisfies condition B. From (C.2) and (C.11), the MISE is:

$$J_n = E\int (\hat{f}_n(x) - f(x))^2 dx$$

$$= \left\{\int \frac{1}{nh2\sqrt{\pi}}\int_z g(x-z;\sqrt{\sigma^2+h^2/2})\,d\alpha(z)\right.$$

$$+ \frac{2}{n^2}\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}\int_{z_1}\int_{z_2} g_2(x-z_1,x-z_2;\beta,\gamma_{m-\ell})\,d\alpha(z_1)\,d\alpha(z_2)$$

$$- 2 \int_{z_1} g(x-z_1;\sigma)\, d\alpha(z_1) \int_{z_2} g(x-z_2;\sqrt{\sigma^2+h^2})\, d\alpha(z_2)$$

$$+ \int_{z_1}\int_{z_2} g(x-z_1;\sigma)\, g(x-z_2;\sigma)\, d\alpha(z_1)\, d\alpha(z_2). \qquad (C.12)$$

Use (B.6) and the equality

$$g_2(x-z_1, x-z_2; \beta, \gamma_{m-\ell}) =$$

$$g\left(x - \frac{(z_1+z_2)}{\sqrt{2}}\,;\, \frac{\beta\sqrt{1+\gamma}}{\sqrt{2}}\right) g\left(z_1-z_2;\beta\sqrt{2(1-\gamma)}\right) \qquad (C.13)$$

to perform the x integrations. We obtain:

$$J_n = \frac{1}{nh2\sqrt{\pi}} \int_z d\alpha(z)$$

$$+ \frac{2}{n^2} \sum_{\ell=1}^{n} \sum_{m=\ell+1}^{n} \int_{z_1}\int_{z_2} g(z_1-z_2;\sqrt{2\sigma^2+h^2})\, d\alpha(z_1)\, d\alpha(z_2)$$

$$+ \int_{z_1}\int_{z_2} g(z_1-z_2;\sqrt{2}\,\sigma)\, d\alpha(z_1)\, d\alpha(z_2). \qquad (C.14)$$

Use Parseval's relation

$$\int_{z_1}\int_{z_2} g(z_1-z_2;\sigma)\, d\alpha(z_1)\, d\alpha(z_2)$$

$$= \frac{1}{2\pi} \int \varphi_\alpha(v)\, \varphi_\alpha^*(v)\, e^{-\frac{1}{2}\sigma^2 v^2}\, dv,$$

where

$$\varphi_\alpha(v) = \int e^{-jvz}\, d\alpha(z),$$

to write

$$2\pi J_n = \frac{\sqrt{\pi}}{nh} + \frac{2}{n^2} \sum_{\ell=1}^{n} \sum_{m=\ell+1}^{n} \int |\varphi_\alpha(v)|^2 e^{-v^2\beta^2(1-\gamma)} dv$$

$$- 2 \int |\varphi_\alpha(v)|^2 e^{-1/2v^2(2\sigma^2+h^2)} dv$$

$$+ \int |\varphi_\alpha(v)|^2 e^{-v^2\sigma^2} dv. \qquad (C.15)$$

Add and subtract

$$(1-\frac{1}{n}) \int |\varphi_\alpha(v)|^2 e^{-v^2(\sigma^2+h^2)} dv.$$

Regrouping terms yields:

$$2\pi J_n = \frac{\sqrt{\pi}}{nh} - \frac{1}{n} \int |\varphi_\alpha(v)|^2 e^{-v^2(\sigma^2+h^2)} dv$$

$$+ \int |\varphi_\alpha(v)|^2 \left\{ e^{-v^2(\sigma^2+h^2)} - 2e^{-\frac{1}{2}v^2(2\sigma^2+h^2)} + e^{-v^2\sigma^2} \right\} dv +$$

$$+ \frac{2}{n^2} \sum_{\ell=1}^{n} \sum_{m=\ell+1}^{n} \int |\varphi_\alpha(v)|^2 \left\{ e^{-v^2\beta^2(1-\gamma)} - e^{-v^2\beta^2} \right\}. \qquad (C.16)$$

Using $|\varphi\alpha(v)| \leq 1$, the second expression on the right hand side is bounded by $\sqrt{\pi}/\sigma$. For the third expression, it follows that as $n \to \infty$, $h(n) \longrightarrow 0$ and

$$\left\{ e^{-v^2(\sigma^2+h^2)} - 2e^{-1/2v^2(2\sigma^2+h^2)} + e^{-v^2\sigma^2} \right\}$$

$$\longrightarrow e^{-v^2\sigma^2} (\frac{5}{8} v^4h^4 + O(h^6)).$$

Hence, as $h \longrightarrow 0$, the third expression of (C.16) can be bounded by $\frac{15}{8}\sqrt{\pi}\sigma^3 h^4$.

The last expression of (C.16) is just slightly more difficult to bound. Substitute for β and $\gamma_{m-\ell}$, and bring the summations inside the integral

$$\frac{2}{n^2}\int e^{-v^2(\sigma^2+h^2)}|\varphi(v)|\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}(e^{+v^2\sigma^2\rho_{m-\ell}}-1)dv. \quad (C.17)$$

Expand the exponential in its power series and let $\tau=m-\ell$. The double summation is then dominated by

$$\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}(e^{+v^2\sigma^2\rho_{m-\ell}}-1)$$

$$=\sum_{\tau=1}^{n}(n-\tau)\sum_{j=1}^{\infty}\frac{(v^2\sigma^2\rho_\tau)^j}{j!}$$

$$=\sum_{j=1}^{\infty}\frac{(+v^2\sigma^2)^j}{j!}\sum_{\tau=1}^{n}(m-\tau)\rho_\tau \leq$$

$$\leq n\sum_{j=1}^{\infty}\frac{(v^2\sigma^2)^j}{j!}\sum_{\tau=1}^{n}\rho_\tau{}^j. \quad (C.18)$$

Under condition B, we have the bound $\sum_{\tau=1}^{\infty}|\rho_\tau|\leq B_2$ (see 2.2.33). Hence, (C.18) is dominated by

$$nB_2\sum_{j=1}^{\infty}\frac{(v^2\sigma^2)^j}{j!} \leq nB_2 e^{+v^2\sigma^2}.$$

Substituting this result into (C.17), we obtain

$$\frac{2}{n^2}\int e^{-v^2(\sigma^2+h^2)}|\varphi_\alpha(v)|^2\sum_{\ell=1}^{n}\sum_{m=\ell+1}^{n}(e^{v^2\sigma^2\rho_{m-\ell}}-1)$$

$$\leq \frac{2B_2}{n}\int |\varphi_\alpha(v)|^2 e^{-v^2h^2}dv$$

$$\leq \frac{2B_2}{n}\frac{\sqrt{\pi}}{h}. \qquad (C.19)$$

Combining all the bounds, we have

$$J_n \leq \frac{1}{2\sqrt{\pi}}\left[\frac{1}{n\sigma}+\frac{1+2B_2}{nh}+\frac{15}{8}\sigma^3h^4\right]. \qquad (C.20)$$

As was the case for the mean-square error (2.3.26), setting $h(n) = n^{-/15}$ gives for the mean integrated square error,

$$J_n = E\int (f_n(x) - f(x))^2dx = O(1/n^{4/5})$$

as $n\to\infty$.

LIST OF REFERENCES

1. Abramson, N., and D. Braverman. "Learning to Recognize Patterns in a Random Environment," IRE Trans. PGIT, Vol. 8, 1962, pp.58-63.

2. Abramson, N., D. Braverman, and G. Sebestyen, "Pattern Recognition and Machine Learning," in Report on Progress in Info. Theory in the U.S.S., 1960 1963, IRE Trans. PGIT, Vol. 9, 1963, pp.257-261.

3. Barrett, J.F., and D.G. Lampard. "An Expansion of Some Second-Order Probability Distributions and Its Applications to Noise Problems, IRE Trans. PGIT, Vol. 1, 1955, pp.10-15.

4. Bilodeau, G.G. "On the Summability of Series of Hermite Polynomials," Journal of Mathematical Analysis and Applications, Vol. 8, 1964, pp. 406-422.

5. Braverman, D. "Learning Filters for Optimum Pattern Recognition," IRE Trans. PGIT, Vol. 8, 1962, pp. 280-285.

6. Brick, D.B., and G. Zames. "Bayes Optimum Filters Derived Using Weiner Canonical Forms," IRE Trans. PGIT, Vol. 8, 1962, pp. 35-46.

7. Brown, J.L. "On a Cross-Correlation Property for Stationary Random Processes," IRE Trans. PGIT, Vol. 3, 1957, pp. 28-31.

8. Capon, J., "On the Asymptotic Effieiency of Locally Optimum Detectors," IRE Trans. PGIT, Vol. 7, 1961, pp. 67-71.

9. Cooper, D.B., and P.W. Cooper. "Nonsupervised Adaptive Signal Detection and Pattern Recognition," Information and Control, Vol. 7, 1964, pp. 416-444.

10. Cramèr, H. Mathematical Methods of Statistics, Princeton University Press, Princeton, New Jersey, 1947.

11. Davenport, W.B., and W.L. Root. An Introduction to the Theory of Random Signals and Noise, McGraw-Hill Book Co., New York, 1958.

12. Erdèlyi, A., W. Magnus, F. Oberhettinger, and F.G. Tricomi. Higher Transcendental Functions, Vol. II, Bateman Manuscript Project, McGraw-Hill Book Co., New York, 1953

13. Erdèlyi, A., W. Magnus, F. Oberhettinger, and F.G. Tricomi. Tables of Integral Transforms, Vol. II, Bateman Manuscript Project, McGraw-Hill Book Co., New York, 1954.

14. Gantmacher, F.R. The Theory of Matrices, Vol. I, trans. by K.A. Hirsch, Chelsea Publishing Co., New York, 1959.

15. Glaser, E.M. "Signal Detection by Adaptive Filters," IRE Trans. PGIT Vol. 7, 1961, pp. 87-98.

16. Halmos, P.R., Lectures on Ergodic Theory, Chelsea Pub. Co., N.Y., 1956.

17. Hancock, J. and E.A. Patrick, Learning Probability Spaces for Classification and Recognition of Patterns With or Without Supervision, Purdue University School of E.E., Electronic Systems Research Lab., Nov. 1965, Tech. Rept. No. TR-EE65-21.

18. Hobson, E.W., The Theory of Functions of a Real Variable, Vol. II, Dover Publications, Inc., N.Y., 1957.

19. Johns, M.V. "An Empirical Bayes Approach to Non-Parametric Two-Way Classification," Studies in Item Analysis and Prediction, Stanford University Press, 1961 ed., H. Solomon

20. Kac, M. "A Note on Learning Signal Detection," IRE Trans. PGIT, Vol. 8, 1962, pp. 126-128.

21. Keehn, D.G., "A Note on Learning for Gaussian Properties," IEEE Trans, PGIT, Vol. 11, 1965, pp. 126-132.

22. Leipnik, R. "Integral Equations, Biorthonormal Expansions and Noise," S.I.A.M., Vol. 7, 1959, pp. 6-30.

23. Lubbock, J. K.,"The Optimization of A Class of Non-Linear-Filters," Institution of Electrical Engineers Proceedings (London), Vol. 107, pt. C, 1960, pp. 60-74.

24. Loève, M., Probability Theory, D. Van Nostrand Co., Inc., Princeton, N.J., 1963, Third Edition.

25. Murthy, V.K., Nonparametric Estimation of Multivariate Densities With Applications, Douglas Missle and Space Systems, Div., Douglas Paper No. 3490.

26. Parzen, E., "On Estimation of a Probability Density and Mode," Ann. Math. Stat. Vol. 33, 1962, 1065-1076.

27. Price, R. and Green, P. E., "A Communication Technique for Multipath Channels," Proc. IRE, Vol. 46, 1958, pp. 555-570.

28. Raviv, J., "Decision Making in Incompletely Known Stochastic Systems," Int. J. Engng. Sci., Vol. 3, 1965, pp. 119-140.

29. Robbins. H., "An Empirical Bayes Approach to Statistics," Proc. Third Berkely Symposium on Statistics and Probability, Vol. I, 1955, pp. 157-164.

30. Robbins. H., "The Empirical Bayes Approach to Testing Statistical Hypotheses," Review of the International Stat, Inst., Vol. 31, pp. 195-208.

31. Robbins, H. "The Empirical Bayes Approach to Statistical Decision Problems," Ann. Math. Stat., Vol. 35, 1964, pp. 1-20.

32. Root, W.L., The Detection of Signals in Gaussian Noise, prepared under NASA Research Grant NsG-2-59, Sept. 1965. To appear as a chapter in a forthcoming book edited by A.V. Balakrishnan.

33. Rosenblatt, M., "Remarks on Some Nonparametric Estimates of a Density Function," Ann. Math. Stat., Vol. 27, 1956, pp. 832-837.

34. Rosenblatt, M., Random Processes, Oxford University Press, N.Y., 1962.

35. Rushford, C. K., Communication in Random or Unknown Channels, Stanford Electronics Lab., 1962, Rept. No. 2004-6.

36. Rushford, C. K., "Adaptive Communication With Sounding Signals in Random Channels," IEEE Int. Conv. Record, 1963, part IV, pp. 102-106.

37. Samuel, E. "An Empirical Bayes Approach to the Testing of Certain Parametric Hypotheses," Ann. Math. Stat., Vol. 34, 1963, pp. 1370-1385.

38. Sansone, G., Orthogonal Functions, Interscience Pub. Inc., N.Y., 1959. Translated from the Italian by A.H. Diamond.

39. Scudder, H.J. III, "Probability of Error of Some Adaptive Pattern-Recognition Machines," IEEE Trans. PGIT, Vol.11, 1965, pp. 363-371.

40. Scudder, H.J. III, "Adaptive Communication Receivers," IEEE Trans. PGIT, Vol. 11, 1965, pp. 167-174.

41. Sebestyen, G., "Pattern Recognition by an Adaptive Process of Sample Set Construction," IRE Trans. PGIT, Vol. 8, 1962, pp. 82-91.

42. Spragins, J., "A Note on the Iterative Application of Bayes Rule," IEEE Trans. PGIT, Vol. 11, 1965, pp. 544-549.

43. Spragins, J., "Learning Without a Teacher," IEEE Trans. PGIT, Vol. 12, 1966, pp. 223-230.

44. Tainiter, M., "Sequential Hypothesis Tests for the r-dependent Marginally Stationary Processes," Ann. Math. Stat., Vol. 37, 1966, pp. 90-97.

45. Watson, G. S. and Leadbetter, M. R., "On the Estimation of the Probability Density, I," Ann. Math. Stat., Vol. 34, 1963, pp. 480-491.

46. Whittle, P. "On the Smoothing of Probability Density Functions," J. Roy. Stat. Soc. Ser. B, Vol. 28, 1958, pp. 334-343.

47. Wong, E. and Thomas, J.B., "On Polynomial Expansions of Second-Order Distributions," SIAM, Vol. 10, 1962, pp. 507-516.