Statistical Signal Processing

Don H. Johnson
Rice University

Contents

Chapter 1

Introduction

MANY signals have a stochastic structure or at least some stochastic component. Some of these signals are a nuisance: noise gets in the way of receiving weak communication signals sent from deep space probes and interference from other wireless calls disturbs cellular telephone systems. Many signals of interest are also stochastic or modeled as such. Compression theory rests on a probabilistic model for every compressed signal. Measurements of physical phenomena, like earthquakes, are stochastic. *Statistical signal processing* algorithms work to extract the good despite the "efforts" of the bad.

This course covers the two basic approaches to statistical signal processing: estimation and detection. In *estimation*, we want to determine a signal's waveform or some signal aspect(s). Typically the parameter or signal we want is buried in noise. Estimation theory shows how to find the best possible — optimal — approach for extracting the information we seek. For example, designing the best filter for removing interference from cell phone calls amounts to a signal waveform estimation algorithm. Determining the delay of a radar signal amounts to a parameter estimation problem. The intent of *detection theory* is to provide rational (instead of arbitrary) techniques for determining which of several conceptions—models—of data generation and measurement is most "consistent" with a given set of data. In digital communication, the received signal must be processed to determine whether it represented a binary "0" or "1"; in radar or sonar, the presence or absence of a target must be determined from measurements of propagating fields; in seismic problems, the presence of oil deposits must be inferred from measurements of sound propagation in the earth. Using detection theory, we will derive signal processing algorithms which will give good answers to questions such as these when the information-bearing signals are corrupted by superfluous signals (noise).

In both areas, we seek *optimal* algorithms: For a given problem statement and optimality criterion, find the approach that minimizes the error. In estimation, our criterion might be mean-squared error or the absolute error. Here, changing the error criterion leads to different estimation algorithms. We have a technical version of the old adage "Beauty is in the eye of the beholder." In detection problems, we might minimize the probability of making an incorrect decision or ensure the detector maximizes the mutual information between input and output. In contrast to estimation, we will find that a single optimal detector minimizes all sensible error criteria. In detection, there is no question what "optimal" means; in estimation, a hundred different papers can be written titled "An optimal estimator" by changing what optimal means. Detection is science; estimation is art.

To solve estimation and/or detection problems, we need to understand stochastic signal models. We begin by reviewing probability theory and stochastic process (random signal) theory. Because we seek to minimize error criteria, we also begin our studies with optimization theory.

Chapter 2

Probability and Stochastic Processes

2.1 Foundations of Probability Theory

2.1.1 Basic Definitions

The basis of probability theory is a set of events—sample space—and a systematic set of numbers—probabilities—assigned to each event. The key aspect of the theory is the system of assigning probabilities. Formally, a *sample space* is the set Ω of all possible outcomes ω_i of an experiment. An *event* is a collection of sample points ω_i determined by some set-algebraic rules governed by the laws of Boolean algebra. Letting A and B denote events, these laws are

$$\begin{aligned} A \cup B &= \{\omega : \omega \in A \text{ or } \omega \in B\} \text{ (union)} \\ A \cap B &= \{\omega : \omega \in A \text{ and } \omega \in B\} \text{ (intersection)} \\ \overline{A} &= \{\omega : \omega \notin A\} \text{ (complement)} \\ \overline{A \cup B} &= \overline{A} \cap \overline{B}. \end{aligned}$$

The null set $\emptyset$ is the complement of Ω. Events are said to be *mutually exclusive* if there is no element common to both events: $A \cap B = \emptyset$.

Associated with each event A_i is a *probability measure* $\Pr[A_i]$, sometimes denoted by π_i, that obeys the *axioms of probability*.

- $\Pr[A_i] \geq 0$
- $\Pr[\Omega] = 1$
- If $A \cap B = \emptyset$, then $\Pr[A \cup B] = \Pr[A] + \Pr[B]$.

The consistent set of probabilities $\Pr[\cdot]$ assigned to events are known as the *a priori probabilities*. From the axioms, probability assignments for Boolean expressions can be computed. For example, simple Boolean manipulations ($A \cup B = A \cup (\overline{A}B)$ and $AB \cup \overline{A}B = B$) lead to

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B].$$

Suppose $\Pr[B] \neq 0$. Suppose we know that the event B has occurred; what is the probability that event A also occurred? This calculation is known as the *conditional probability* of A given B and is denoted by $\Pr[A|B]$. To evaluate conditional probabilities, consider B to be the sample space rather than Ω. To obtain a probability assignment under these circumstances consistent with the axioms of probability, we must have

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

The event is said to be *statistically independent* of B if $\Pr[A|B] = \Pr[A]$: the occurrence of the event B does not change the probability that A occurred. When independent, the probability of their intersection $\Pr[A \cap B]$ is given by the product of the *a priori* probabilities $\Pr[A] \cdot \Pr[B]$. This property is necessary and sufficient for the independence of the two events. As $\Pr[A|B] = \Pr[A \cap B] / \Pr[B]$ and $\Pr[B|A] = \Pr[A \cap B] / \Pr[A]$, we obtain *Bayes' Rule*.

$$\Pr[B|A] = \frac{\Pr[A|B] \cdot \Pr[B]}{\Pr[A]}$$

2.1.2 Random Variables and Probability Density Functions

A *random variable* X is the assignment of a number—real or complex—to each sample point in sample space; mathematically, $X : \Omega \mapsto \mathbb{R}$. Thus, a random variable can be considered a function whose domain is a set and whose range are, most commonly, a subset of the real line. This range could be discrete-valued (especially when the domain Ω is discrete). In this case, the random variable is said to be *symbolic-valued*. In some cases, the symbols can be related to the integers, and then the values of the random variable can be ordered. When the range is continuous, an interval on the real-line say, we have a *continuous-valued* random variable. In some cases, the random variable is a *mixed* random variable: it is both discrete- and continuous-valued.

The *probability distribution function* or *cumulative* can be defined for continuous, discrete (only if an ordering exists), and mixed random variables.

$$P_X(x) \equiv \Pr[X \leq x] .$$

Note that X denotes the random variable and x denotes the argument of the distribution function. Probability distribution functions are increasing functions: if $A = \{\omega : X(\omega) \leq x_1\}$ and $B = \{\omega : x_1 < X(\omega) \leq x_2\}$, $\Pr[A \cup B] = \Pr[A] + \Pr[B] \implies P_X(x_2) = P_X(x_1) + \Pr[x_1 < X \leq x_2]$,* which means that $P_X(x_2) \geq P_X(x_1)$, $x_1 \leq x_2$.

The *probability density function* $p_X(x)$ is defined to be that function when integrated yields the distribution function.

$$P_X(x) = \int_{-\infty}^{x} p_X(\alpha)\, d\alpha$$

As distribution functions may be discontinuous when the random variable is discrete or mixed, we allow density functions to contain impulses. Furthermore, density functions must be non-negative since their integrals are increasing.

2.1.3 Function of a Random Variable

When random variables are real-valued, we can consider applying a real-valued function. Let $Y = f(X)$; in essence, we have the sequence of maps $f : \Omega \mapsto \mathbb{R} \mapsto \mathbb{R}$, which is equivalent to a simple mapping from sample space Ω to the real line. Mappings of this sort constitute the definition of a random variable, leading us to conclude that Y is a random variable. Now the question becomes "What are Y's probabilistic properties?". The key to determining the probability density function, which would allow calculation of the mean and variance, for example, is to use the probability *distribution* function.

For the moment, assume that $f(\cdot)$ is a strictly-monotonically increasing function. The probability distribution of Y we seek is

$$\begin{aligned} P_Y(y) &= \Pr[Y \leq y] \\ &= \Pr[f(X) \leq y] \\ &= \Pr[X \leq f^{-1}(y)] \qquad (\dagger) \\ &= P_X\left(f^{-1}(y)\right) \end{aligned}$$

Equation (†) is the key step; here, $f^{-1}(y)$ is the inverse function. Because $f(\cdot)$ is a strictly increasing function, the underlying portion of sample space corresponding to $Y \leq y$ must be the same as that corresponding to

*What property do the sets A and B have that makes this expression correct?

$X \leq f^{-1}(y)$. We can find Y's density by evaluating the derivative.

$$p_y(y) = \frac{df^{-1}(y)}{dy} p_X\left(f^{-1}(y)\right)$$

The derivative term amounts to $1/f'(x)|_{x=f^{-1}(y)}$.

The style of this derivation applies to monotonically decreasing functions as well. The difference is that the set corresponding to $Y \leq y$ now corresponds to $X \geq f^{-1}(y)$. Now, $P_Y(y) = 1 - P_X\left(f^{-1}(y)\right)$. The probability density function of a monotonic — increasing or decreasing — function of a random variable is found according to the formula

$$\boxed{p_y(y) = \left|\frac{1}{f'\left(f^{-1}(y)\right)}\right| p_X\left(f^{-1}(y)\right)}.$$

Example

Suppose X has an exponential probability density: $p_X(x) = e^{-x}\mathrm{u}(x)$, where $\mathrm{u}(x)$ is the unit-step function. We have $Y = X^2$. Because the square-function is monotonic over the positive real line, our formula applies. We find that

$$p_Y(y) = \frac{1}{2\sqrt{y}} e^{-\sqrt{y}},\ y > 0\ .$$

Although difficult to show, this density indeed integrates to one.

2.1.4 Expected Values

The *expected value* of a function $f(\cdot)$ of a random variable X is defined to be

$$\mathsf{E}[f(X)] = \int_{-\infty}^{\infty} f(x) p_X(x)\, dx.$$

Several important quantities are expected values, with specific forms for the function $f(\cdot)$.

- $f(X) = X$.
  The *expected value* or *mean* of a random variable is the center-of-mass of the probability density function. We shall often denote the expected value by m_X or just m when the meaning is clear. Note that the expected value can be a number never assumed by the random variable ($p_X(m)$ can be zero). An important property of the expected value of a random variable is *linearity*: $\mathsf{E}[aX] = a\,\mathsf{E}[X]$, a being a scalar.
- $f(X) = X^2$.
  $\mathsf{E}[X^2]$ is known as the *mean squared value* of X and represents the "power" in the random variable.
- $f(X) = (X - m_X)^2$.
  The so-called second central difference of a random variable is its *variance*, usually denoted by σ_X^2. This expression for the variance simplifies to $\sigma_X^2 = \mathsf{E}[X^2] - \mathsf{E}^2[X]$, which expresses the variance operator $\mathsf{var}[\cdot]$. The square root of the variance σ_X is the *standard deviation* and measures the spread of the distribution of X. Among all possible second differences $(X - c)^2$, the minimum value occurs when $c = m_X$ (simply evaluate the derivative with respect to c and equate it to zero).
- $f(X) = X^n$.
  $\mathsf{E}[X^n]$ is the n^{th} *moment* of the random variable and $\mathsf{E}\left[(X - m_X)^n\right]$ the n^{th} central moment.
- $f(X) = e^{jvX}$ and $f(X) = e^{sX}$.
  The *characteristic function* of a random variable is essentially the Fourier Transform of the probability density function. Note that the sign of the complex exponential is positive; this choice is a long-standing convention in probability.

$$\mathsf{E}\left[e^{jvX}\right] \equiv \Phi_X(jv) = \int_{-\infty}^{\infty} p_X(x) e^{+jvx}\, dx$$

The moments of a random variable can be calculated from the derivatives of the characteristic function evaluated at the origin.

$$\mathsf{E}[X^n] = j^{-n}\frac{d^n\Phi_X(j\nu)}{d\nu^n}\bigg|_{\nu=0}$$

We can also define the *moment generating function* as the Laplace transform of the density.

$$\Phi_X(s) \equiv \mathsf{E}\left[e^{sX}\right] = \int_{-\infty}^{\infty} p_X(x)e^{+sx}\,dx$$

With this definition, the relationship of the derivatives of the moment generating function with moments is more direct.

$$\mathsf{E}[X^n] = \frac{d^n\Phi_X(s)}{ds^n}\bigg|_{s=0}$$

If the more conventional minus sign were chosen, this expression would be a bit more complicated.

2.1.5 Jointly Distributed Random Variables

Two (or more) random variables can be defined over the same sample space: $X:\Omega\mapsto\mathbb{R}$, $Y:\Omega\mapsto\mathbb{R}$. More generally, we can have a random vector (dimension N) $\mathbf{X}:\Omega\mapsto\mathbb{R}^N$. First, let's consider the two-dimensional case: $\mathbf{X}=\{X,Y\}$. Just as with jointly defined events, the *joint distribution function* is easily defined.

$$P_{X,Y}(x,y) \equiv \Pr[\{X\le x\}\cap\{Y\le y\}]$$

The *joint probability density function* $p_{X,Y}(x,y)$ is related to the distribution function via double integration.

$$P_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} p_{X,Y}(\alpha,\beta)\,d\alpha\,d\beta \quad \text{or} \quad p_{X,Y}(x,y) = \frac{\partial^2 P_{X,Y}(x,y)}{\partial x\partial y}$$

Since $\lim_{y\to\infty}P_{X,Y}(x,y)=P_X(x)$, the so-called *marginal density functions* can be related to the joint density function.

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x,\beta)\,d\beta \text{ and } p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(\alpha,y)\,d\alpha$$

Extending the ideas of conditional probabilities, the *conditional probability density function* $p_{X|Y}(x|Y=y)$ is defined (when $p_Y(y)\neq 0$) as

$$p_{X|Y}(x|Y=y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Two random variables are *statistically independent* when $p_{X|Y}(x|Y=y)=p_X(x)$, which is equivalent to the condition that the joint density function is separable: $p_{X,Y}(x,y)=p_X(x)\cdot p_Y(y)$.

For jointly defined random variables, expected values are defined similarly as with single random variables. Probably the most important joint moment is the *covariance*:

$$\mathsf{cov}[X,Y] \equiv \mathsf{E}\left[(X-m_X)(Y-m_Y)\right] = \mathsf{E}[XY]-\mathsf{E}[X]\cdot\mathsf{E}[Y], \quad \text{where } \mathsf{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,p_{X,Y}(x,y)\,dx\,dy\,.$$

Related to the covariance is the (confusingly named) *correlation coefficient*: the covariance normalized by the standard deviations of the component random variables.

$$\rho_{X,Y} = \frac{\mathsf{cov}[X,Y]}{\sigma_X\sigma_Y}$$

When two random variables are *uncorrelated*, their covariance and correlation coefficient equals zero so that $\mathsf{E}[XY]=\mathsf{E}[X]\,\mathsf{E}[Y]$. Statistically independent random variables are always uncorrelated, but uncorrelated random variables can be dependent.*

A *conditional expected value* is the mean of the conditional density.

$$\mathsf{E}[X|Y=y] = \int_{-\infty}^{\infty} x\,p_{X|Y}(x|Y=y)\,dx$$

*Let X be uniformly distributed over $[-1,1]$ and let $Y=X^2$. The two random variables are uncorrelated, but are clearly not independent.

Note that the conditional expected value is now a function of Y and is therefore a random variable. Consequently, it too has an expected value, which is easily evaluated to be the expected value of X.

$$\mathsf{E}\big[\mathsf{E}[X|Y]\big] = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} x p_{X|Y}(x|Y=y)\,dx\right] p_Y(y)\,dy = \mathsf{E}[X]$$

More generally, the expected value of a function of two random variables can be shown to be the expected value of a conditional expected value: $\mathsf{E}\big[f(X,Y)\big] = \mathsf{E}\big[\mathsf{E}[f(X,Y)|Y]\big]$. This kind of calculation is frequently simpler to evaluate than trying to find the expected value of $f(X,Y)$ "all at once." A particularly interesting example of this simplicity is the *random sum of random variables*. Let L be a random variable and $\{X_l\}$ a sequence of random variables. We will find occasion to consider the quantity $\sum_{l=1}^{L} X_l$. Assuming that the each component of the sequence has the same expected value $\mathsf{E}[X]$, the expected value of the sum is found to be

$$\begin{aligned}\mathsf{E}[S_L] &= \mathsf{E}\left[\mathsf{E}\left[\textstyle\sum_{l=1}^{L} X_l | L\right]\right]\\ &= \mathsf{E}\big[L\cdot\mathsf{E}[X]\big]\\ &= \mathsf{E}[L]\cdot\mathsf{E}[X]\end{aligned}$$

2.1.6 Random Vectors

A *random vector* $\mathbf{X}$ is an ordered sequence of random variables $\mathbf{X} = \mathrm{col}[X_1,\ldots,X_L]$. The density function of a random vector is defined in a manner similar to that for pairs of random variables. The expected value of a random vector is the vector of expected values.

$$\mathsf{E}[\mathbf{X}] = \int_{-\infty}^{\infty} \mathbf{x} p_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x} = \mathrm{col}\big[\mathsf{E}[X_1],\ldots,\mathsf{E}[X_L]\big]$$

The *covariance matrix* $\mathbf{K}_X$ is an $L\times L$ matrix consisting of all possible covariances among the random vector's components.

$$\mathbf{K}_{ij}^{X} = \mathrm{cov}[X_i,X_j] = \mathsf{E}[X_i X_j^*] - \mathsf{E}[X_i]\,\mathsf{E}[X_j^*] \quad i,j = 1,\ldots,L$$

Using matrix notation, the covariance matrix can be written as $\mathbf{K}_X = \mathsf{E}\big[(\mathbf{X}-\mathsf{E}[\mathbf{X}])(\mathbf{X}-\mathsf{E}[\mathbf{X}])'\big]$. Using this expression, the covariance matrix is seen to be a symmetric matrix and, when the random vector has no zero-variance component, its covariance matrix is positive-definite. Note in particular that when the random variables are real-valued, the diagonal elements of a covariance matrix equal the variances of the components: $\mathbf{K}_{ii}^{X} = \sigma_{X_i}^2$. *Circular* random vectors are complex-valued with uncorrelated, identically distributed, real and imaginary parts. In this case, $\mathsf{E}\left[|X_i|^2\right] = 2\sigma_{X_i}^2$ and $\mathsf{E}\left[X_i^2\right] = 0$. By convention, $\sigma_{X_i}^2$ denotes the variance of the real (or imaginary) part. The characteristic function of a real-valued random vector is defined to be

$$\Phi_{\mathbf{X}}(j\boldsymbol{\nu}) = \mathsf{E}\left[e^{j\boldsymbol{\nu}'\mathbf{X}}\right].$$

2.1.7 Single function of a random vector

Just as shown in §2.1.3, the key tool is the distribution function. When $Y = f(\mathbf{X})$, a scalar-valued function of a vector, we need to find that portion of the domain that corresponds to $f(\mathbf{X}) \le y$. Once this region is determined, the density can be found.

For example, the maximum of a random vector is a random variable whose probability density is usually quite different than the distributions of the vector's components. The probability that the maximum is less than some number μ is equal to the probability that *all* of the components are less than μ.

$$\Pr[\max\mathbf{X} < \mu] = P_{\mathbf{X}}(\mu,\ldots,\mu)$$

Assuming that the components of $\mathbf{X}$ are statistically independent, this expression becomes

$$\Pr[\max\mathbf{X} < \mu] = \prod_{i=1}^{\dim\mathbf{X}} P_{X_i}(\mu),$$

and the density of the maximum has an interesting answer.

$$p_{\max \mathbf{X}}(\mu) = \sum_{j=1}^{\dim \mathbf{X}} p_{X_j}(\mu) \prod_{i \neq j} P_{X_i}(\mu)$$

When the random vector's components are identically distributed, we have

$$p_{\max \mathbf{X}}(\mu) = (\dim \mathbf{X}) p_X(\mu) P_X^{(\dim \mathbf{X})-1}(\mu) \, .$$

2.1.8 Several functions of a random vector

When we have a vector-valued function of a vector (and the input and output dimensions don't necessarily match), finding the joint density of the function can be quite complicated, but the recipe of using the joint distribution function still applies. In some (intersting) cases, the derivation flows nicely. Consider the case where $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A}$ is an invertible matrix.

$$\begin{aligned} P_{\mathbf{Y}}(\mathbf{y}) &= \Pr[\mathbf{AX} \leq \mathbf{y}] \\ &= \Pr\left[\mathbf{X} \leq \mathbf{A}^{-1}\mathbf{y}\right] \\ &= P_{\mathbf{X}}\left(\mathbf{A}^{-1}\mathbf{y}\right) \end{aligned}$$

To find the density, we need to evaluate the N^{th}-order mixed derivative (N is the dimension of the random vectors). The Jacobian appears and in this case, the Jacobian is the determinant of the matrix $\mathbf{A}$.

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|} p_{\mathbf{X}}\left(\mathbf{A}^{-1}\mathbf{y}\right) \tag{2.1}$$

2.1.9 The Gaussian Random Variable

The random variable X is said to be a *Gaussian random variable** if its probability density function has the form

$$\boxed{p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}}.$$

The mean of such a Gaussian random variable is m and its variance σ^2. As a shorthand notation, this information is denoted by $x \sim \mathcal{N}(m, \sigma^2)$. The characteristic function $\Phi_X(\cdot)$ of a Gaussian random variable is given by

$$\Phi_X(j\nu) = e^{jm\nu} \cdot e^{-\sigma^2\nu^2/2} \, .$$

No closed form expression exists for the probability distribution function of a Gaussian random variable. For a zero-mean, unit-variance, Gaussian random variable $\left(\mathcal{N}(0,1)\right)$, the probability that it *exceeds* the value x is denoted by $Q(x)$.

$$\Pr[X > x] = 1 - P_X(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\alpha^2/2} \, d\alpha \equiv Q(x)$$

A plot of $Q(\cdot)$ is shown in Fig. 2.1. When the Gaussian random variable has non-zero mean and/or non-unit variance, the probability of it exceeding x can also be expressed in terms of $Q(\cdot)$.

$$\Pr[X > x] = Q\left(\frac{x-m}{\sigma}\right), \quad X \sim \mathcal{N}(m, \sigma^2)$$

Integrating by parts, $Q(\cdot)$ is bounded (for $x > 0$) by

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{x}{1+x^2} e^{-x^2/2} \leq Q(x) \leq \frac{1}{\sqrt{2\pi}x} e^{-x^2/2} \, . \tag{2.2}$$

As x becomes large, these bounds approach each other and either can serve as an approximation to $Q(\cdot)$; the upper bound is usually chosen because of its relative simplicity. The lower bound can be improved; noting

*Gaussian random variables are also known as *normal* random variables.

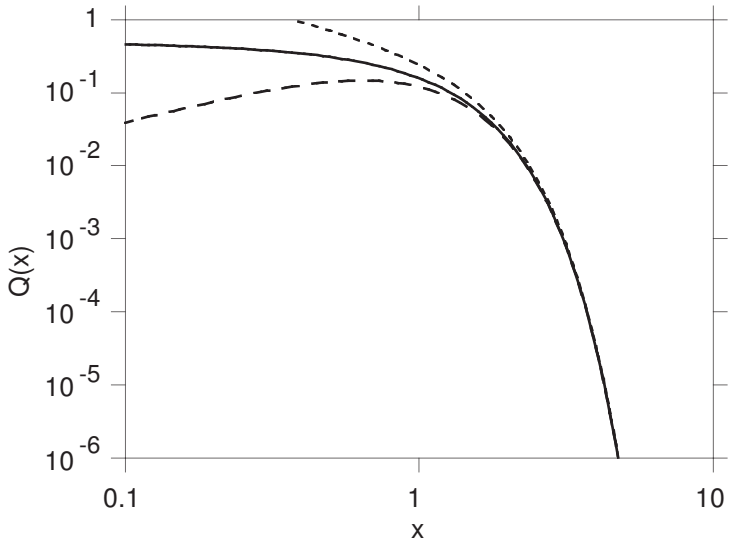


Figure 2.1: The function $Q(\cdot)$ is plotted on logarithmic coordinates. Beyond values of about two, this function decreases quite rapidly. Two approximations are also shown that correspond to the upper and lower bounds given by Eq. 2.2.

that the term $x/(1+x^2)$ decreases for $x<1$ and that $Q(x)$ increases as x decreases, the term can be replaced by its value at $x=1$ without affecting the sense of the bound for $x \leq 1$.

$$\frac{1}{2\sqrt{2\pi}} e^{-x^2/2} \leq Q(x), \quad x \leq 1 \tag{2.3}$$

We will have occasion to evaluate the expected value of $\exp\{aX+bX^2\}$ where $X \sim \mathcal{N}(m,\sigma^2)$ and a, b are constants. By definition,

$$\mathsf{E}[e^{aX+bX^2}] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\{ax+bx^2-(x-m)^2/(2\sigma^2)\}\,dx$$

The argument of the exponential requires manipulation (*i.e.*, completing the square) before the integral can be evaluated. This expression can be written as

$$-\frac{1}{2\sigma^2}\{(1-2b\sigma^2)x^2-2(m+a\sigma^2)x+m^2\}\,.$$

Completing the square, this expression can be written

$$-\frac{1-2b\sigma^2}{2\sigma^2}\left(x-\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2+\frac{1-2b\sigma^2}{2\sigma^2}\left(\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2-\frac{m^2}{2\sigma^2}$$

We are now ready to evaluate the integral. Using this expression,

$$\mathsf{E}[e^{aX+bX^2}] = \exp\left\{\frac{1-2b\sigma^2}{2\sigma^2}\left(\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2-\frac{m^2}{2\sigma^2}\right\}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1-2b\sigma^2}{2\sigma^2}\left(x-\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2\right\}dx\,.$$

Let

$$\alpha = \frac{x-\frac{m+a\sigma^2}{1-2b\sigma^2}}{\frac{\sigma}{\sqrt{1-2b\sigma^2}}},$$

which implies that we must require that $1-2b\sigma^2>0$ $\big($or $b<1/(2\sigma^2)\big)$. We then obtain

$$\mathsf{E}\left[e^{aX+bX^2}\right] = \exp\left\{\frac{1-2b\sigma^2}{2\sigma^2}\left(\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2-\frac{m^2}{2\sigma^2}\right\}\frac{1}{\sqrt{1-2b\sigma^2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{\alpha^2}{2}}\,d\alpha\,.$$

The integral equals unity, leaving the result

$$\boxed{\mathsf{E}[e^{aX+bX^2}] = \frac{\exp\left\{\frac{1-2b\sigma^2}{2\sigma^2}\left(\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2 - \frac{m^2}{2\sigma^2}\right\}}{\sqrt{1-2b\sigma^2}}, b < \frac{1}{2\sigma^2}}$$

Important special cases are

1. $a = 0, X \sim \mathcal{N}(m, \sigma^2)$.
$$\mathsf{E}[e^{bX^2}] = \frac{\exp\left\{\frac{bm^2}{1-2b\sigma^2}\right\}}{\sqrt{1-2b\sigma^2}}$$
2. $a = 0, X \sim \mathcal{N}(0, \sigma^2)$.
$$\mathsf{E}[e^{bX^2}] = \frac{1}{\sqrt{1-2b\sigma^2}}$$
3. $X \sim \mathcal{N}(0, \sigma^2)$.
$$\mathsf{E}[e^{aX+bX^2}] = \frac{\exp\left\{\frac{a^2\sigma^2}{2(1-2b\sigma^2)}\right\}}{1-2b\sigma^2}$$

The real-valued random vector $\mathbf{X}$ is said to be a *Gaussian random vector* if its joint distribution function has the form

$$\boxed{p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det[2\pi\mathbf{K}]}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\right\}}.$$

If complex-valued, the joint distribution of a circular Gaussian random vector is given by

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det[\pi\mathbf{K}]}} \exp\left\{-(\mathbf{x}-\mathbf{m}_X)'\mathbf{K}_X^{-1}(\mathbf{x}-\mathbf{m}_X)\right\}. \tag{2.4}$$

The vector $\mathbf{m}_X$ denotes the expected value of the Gaussian random vector and $\mathbf{K}_X$ its covariance matrix.

$$\mathbf{m}_X = \mathsf{E}[\mathbf{X}] \qquad \mathbf{K}_X = \mathsf{E}[\mathbf{X}\mathbf{X}'] - \mathbf{m}_X\mathbf{m}_X'$$

As in the univariate case, the Gaussian distribution of a random vector is denoted by $\mathbf{X} \sim \mathcal{N}(\mathbf{m}_X, \mathbf{K}_X)$. After applying a linear transformation to Gaussian random vector, such as $\mathbf{Y} = \mathbf{A}\mathbf{X}$, the result is also a Gaussian random vector (a random variable if the matrix is a row vector): $\mathbf{Y} \sim \mathcal{N}(\mathbf{A}\mathbf{m}_X, \mathbf{A}\mathbf{K}_X\mathbf{A}')$ [see equation (2.1) {8}]. The characteristic function of a Gaussian random vector is given by

$$\Phi_{\mathbf{X}}(j\boldsymbol{\nu}) = \exp\left\{+j\boldsymbol{\nu}^t\mathbf{m}_X - \frac{1}{2}\boldsymbol{\nu}^t\mathbf{K}_X\boldsymbol{\nu}\right\}.$$

From this formula, the N^{th}-order moment formula for jointly distributed Gaussian random variables is easily derived.*

$$\boxed{\mathsf{E}[X_1 \cdots X_N] = \begin{cases} \sum_{\text{all}\mathscr{P}_N} \mathsf{E}[X_{\mathscr{P}_N(1)}X_{\mathscr{P}_N(2)}] \cdots \mathsf{E}[X_{\mathscr{P}_N(N-1)}X_{\mathscr{P}_N(N)}], & N \text{ even} \\ \sum_{\text{all}\mathscr{P}_N} \mathsf{E}[X_{\mathscr{P}_N(1)}] \mathsf{E}[X_{\mathscr{P}_N(2)}X_{\mathscr{P}_N(3)}] \cdots \mathsf{E}[X_{\mathscr{P}_N(N-1)}X_{\mathscr{P}_N(N)}], & N \text{ odd} \end{cases}},$$

where $\mathscr{P}_N$ denotes a permutation of the first N integers and $\mathscr{P}_N(i)$ the i^{th} element of the permutation. For example, we obtain the following important formula for the fourth joint moment

$$\mathsf{E}[X_1X_2X_3X_4] = \mathsf{E}[X_1X_2]\,\mathsf{E}[X_3X_4] + \mathsf{E}[X_1X_3]\,\mathsf{E}[X_2X_4] + \mathsf{E}[X_1X_4]\,\mathsf{E}[X_2X_3],$$

which, in the important special case that the random variables equal each other ($X_1 = X_2 = X_3 = X_4$), becomes $\mathsf{E}[X^4] = 3\big(\mathsf{E}[X^2]\big)^2$. In the special case where the Gaussian random variable has mean zero,

$$\boxed{\mathsf{E}[X^4] = 3\sigma_X^4}.$$

Note that this result was derived for Gaussian random variables; it usually does *not* apply to non-Gaussian random variables.

* $\mathsf{E}[X_1 \cdots X_N] = j^{-N} \frac{\partial^N}{\partial\nu_1 \cdots \partial\nu_N} \Phi_{\mathbf{X}}(j\boldsymbol{\nu})\Big|_{\boldsymbol{\nu}=\mathbf{0}}$.

2.1.10 The Central Limit Theorem

Let $\{X_l\}$ denote a sequence of independent, identically distributed, random variables. Assuming they have zero means and finite variances (equaling σ^2), the Central Limit Theorem states that the sum $\sum_{l=1}^{L} X_l/\sqrt{L}$ converges in distribution to a Gaussian random variable.

$$\frac{1}{\sqrt{L}}\sum_{l=1}^{L} X_l \xrightarrow{L\to\infty} \mathcal{N}(0,\sigma^2)$$

Because of its generality, this theorem is often used to simplify calculations involving *finite* sums of non-Gaussian random variables. However, attention is seldom paid to the *convergence rate* of the Central Limit Theorem. Kolmogorov, the famous twentieth century mathematician, is reputed to have said "The Central Limit Theorem is a dangerous tool in the hands of amateurs." Let's see what he meant.

Taking $\sigma^2 = 1$, the key result is that the magnitude of the difference between $P(x)$, defined to be the probability that the sum given above exceeds x, and $Q(x)$, the probability that a unit-variance Gaussian random variable exceeds x, is bounded by a quantity inversely related to the square root of L [26: Theorem 24].

$$|P(x) - Q(x)| \le c \cdot \frac{\mathsf{E}\left[|X|^3\right]}{\sigma^3} \cdot \frac{1}{\sqrt{L}}$$

The constant of proportionality c is a number known to be about 0.8 [43: p. 6]. The ratio of absolute third moment of X_l to the cube of its standard deviation, known as the skew and denoted by γ_X, depends only on the distribution of X_l and is independent of scale. This bound on the absolute error has been shown to be tight [26: pp. 79ff]. Using our lower bound for $Q(\cdot)$ (Eq. 2.3 {9}), we find that the relative error in the Central Limit Theorem approximation to the distribution of finite sums is bounded for $x > 0$ as

$$\boxed{\frac{|P(x) - Q(x)|}{Q(x)} \le c\gamma_X \sqrt{\frac{2\pi}{L}} e^{+x^2/2} \cdot \begin{cases} 2, & x \le 1 \\ \frac{1+x^2}{x}, & x > 1 \end{cases}}.$$

Suppose we require that the relative error not exceed some specified value ε. The normalized (by the standard deviation) boundary x at which the approximation is evaluated must not violate

$$\frac{L\varepsilon^2}{2\pi c^2 \gamma_X^2} \ge e^{x^2} \cdot \begin{cases} 4 & x \le 1 \\ \left(\frac{1+x^2}{x}\right)^2 & x > 1 \end{cases}.$$

As shown in Fig. 2.2, the right side of this equation is a monotonically increasing function.

For example, if $\varepsilon = 0.1$ and taking $c\gamma_X$ arbitrarily to be unity (a reasonable value), the upper limit of the preceding equation becomes $1.6 \times 10^{-3} L$. Examining Fig. 2.2, we find that for $L = 10{,}000$, x must not exceed 1.17. Because we have normalized to unit variance, this example suggests that the Gaussian approximates the distribution of a ten-thousand term sum only over a range corresponding to an 76% area about the mean. Consequently, the Central Limit Theorem, as a finite-sample distributional approximation, is only guaranteed to hold near the mode of the Gaussian, with *huge* numbers of observations needed to specify the tail behavior. Realizing this fact will keep us from being ignorant amateurs.

2.2 Stochastic Processes

2.2.1 Basic Definitions

A *random* or *stochastic* process is the assignment of a function of a real variable to each sample point ω in sample space. Thus, the process $X(\omega,t)$ can be considered a function of two variables. For each ω, the time function must be well-behaved and may or may not look random to the eye. Each time function of the process is called a *sample function* and must be defined over the entire domain of interest. For each t, we have a function of ω, which is precisely the definition of a random variable. Hence the *amplitude* of a random process is a random variable. The *amplitude distribution* of a process refers to the probability

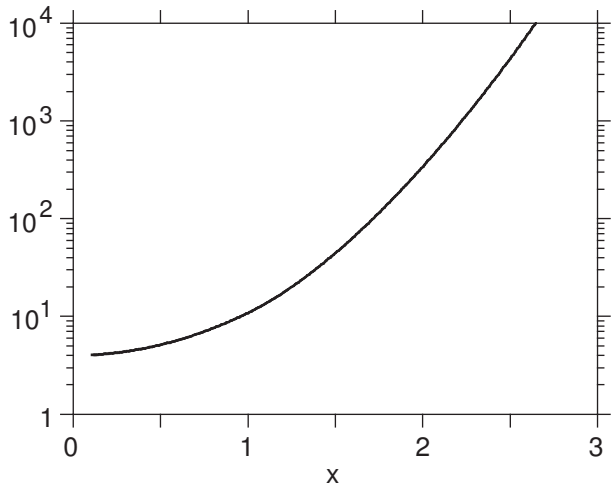


Figure 2.2: The quantity which governs the limits of validity for numerically applying the Central Limit Theorem on finite numbers of data is shown over a portion of its range. To judge these limits, we must compute the quantity $L\varepsilon^2/2\pi c^2\gamma_X$, where ε denotes the desired percentage error in the Central Limit Theorem approximation and L the number of observations. Selecting this value on the vertical axis and determining the value of x yielding it, we find the normalized ($x = 1$ implies unit variance) upper limit on an L-term sum to which the Central Limit Theorem is guaranteed to apply. Note how rapidly the curve increases, suggesting that large amounts of data are needed for accurate approximation.

density function of the amplitude: $p_{X(t)}(x)$. By examining the process's amplitude at several instants, the joint amplitude distribution can also be defined. A process is completely *specified* if the joint probability distribution of the process's amplitudes at an arbitrary number of times and arbitrary time instances can be found. For our purposes, a process is said to be *stationary* when the joint amplitude distribution remains the same when the time instances are shifted by the same delay.

$$p_{X(t_1-\Delta),\ldots,X(t_N-\Delta)}(x_1,\ldots,x_N) = p_{X(t_1),\ldots,X(t_N)}(x_1,\ldots,x_N),\ \forall\Delta,N$$

The *expected value* or *mean* of a process is the expected value of the amplitude at each t.

$$\mathsf{E}[X(t)] = m_X(t) = \int_{-\infty}^{\infty} x p_{X(t)}(x)\,dx$$

For the most part, we take the mean to be zero. The *correlation function* is the first-order joint moment between the process's amplitudes at two times.

$$R_X(t_1,t_2) = \mathsf{E}[X(t_1)X(t_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 p_{X(t_1),X(t_2)}(x_1,x_2)\,dx_1\,dx_2$$

For the bivariate case, the joint distribution for stationary processes depends *only* on difference between the time instances (let $N =$ and $\Delta = t_1$ in the above equation. Consequently, correlation functions of stationary processes depend only on $|t_1 - t_2|$. In this case, correlation functions are really functions of a single variable (the time difference) and are usually written as $R_X(\tau)$ where $\tau = t_1 - t_2$. In the special case when the mean is a constant and the correlation function is a function of τ, the process is said to be *wide-sense stationary*. Many processes can be wide-sense stationary but not stationary in the strict sense. As described subsequently, many properties of random processes can be found using only the process's mean and correlation function. Having these constitutes the *second-order specification* of the process. Clearly, having a second-order specification does *not* specify the process completely: you can't derive the joint probability distribution from the mean and correlation. Related to the correlation function is the *covariance function* $K_X(\tau)$, which equals the correlation function minus the square of the mean.

$$K_X(\tau) = R_X(\tau) - m_X^2$$

The variance of the process equals the covariance function evaluated as the origin. Given any two of the mean, correlation function and covariance function, the third can be found, even when the process is non-stationary.

The *power spectrum* of a wide-sense stationary process is the Fourier Transform of the correlation function.

$$\mathscr{S}_X(f) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau}\, d\tau$$

When a stationary process $X(t)$ is passed through a stable linear, time-invariant filter, the resulting output $Y(t)$ is also a stationary process having power density spectrum

$$\boxed{\mathscr{S}_Y(f) = |H(f)|^2 \mathscr{S}_X(f)},$$

where $H(f)$ is the filter's transfer function. This result is easily derived by expressing the output process as the convolution of the input process with the filter's impulse response, finding the correlation function of the output and calculating the Fourier transform of the result.

$$\begin{aligned}
\mathsf{E}[Y(t)Y(u)] &= \mathsf{E}\left[\int_{-\infty}^{\infty} h(t-\alpha)X(\alpha)\,d\alpha \cdot \int_{-\infty}^{\infty} h(u-\beta)X(\beta)\,d\beta\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\alpha)h(u-\beta)\,\mathsf{E}[X(\alpha)X(\beta)]\,d\alpha d\beta \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\alpha)h(u-\beta)R_x(\alpha-\beta)\,d\alpha d\beta \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-\xi-\beta)h(u-\beta)R_x(\xi)\,d\xi d\beta \quad [\xi = \alpha-\beta] \\
R_Y(\tau) &= h(\tau)\otimes h(-\tau)\otimes R_x(\tau) \\
\Longrightarrow\ \mathscr{S}_Y(f) &= |H(f)|^2 \mathscr{S}_X(f)
\end{aligned}$$

A particularly important example of a random process is *white noise*. The process $X(t)$ is said to be white if it has zero mean and a correlation function proportional to an impulse.

$$\mathsf{E}\big[X(t)\big] = 0 \quad R_X(\tau) = \frac{N_0}{2}\delta(\tau)$$

The power spectrum of white noise is constant for all frequencies, equaling $N_0/2$. which is known as the *spectral height*.* Because the spectrum is constant, the output of a linear, time-invariant system having a white noise input has a power spectrum that is proportional to the magnitude-squared of the system's transfer function.

$$\mathscr{S}_Y(f) = \frac{N_0}{2}\cdot|H(f)|^2$$

You can generate random processes having a specific power spectrum by appropriately choosing the linear system's transfer function and passing white noise through it. By the way, random processes generated by passing white noise through a linear filter are said to be *linear processes*.

If the filter is *not* stable, interesting linear processes can arise. The most important case occurs when an integrator has white noise as its input.

$$Y(t) = \int_0^t X(\alpha)\,d\alpha, \quad t > 0$$

Clearly, the mean of the output is zero since the input has zero mean. More interesting is the correlation

*The curious reader can track down why the spectral height of white noise has the fraction one-half in it. This definition is the convention.

function.

$$\begin{aligned} R_Y(t,u) &= \mathsf{E}\left[\int_0^t X(\alpha)\,d\alpha \int_0^u X(\beta)\,d\beta\right], \quad t,u>0 \\ &= \int_0^t \int_0^u \mathsf{E}[X(\alpha)X(\beta)]\,d\alpha\,d\beta \\ &= \int_0^t \int_0^u \frac{N_0}{2}\delta(\alpha-\beta)\,d\alpha\,d\beta \end{aligned}$$

Performing the "α" integral first, we have

$$\begin{aligned} &= \frac{N_0}{2}\int_0^u \mathrm{u}(t-\beta)\,d\beta \qquad \mathrm{u}(\cdot) \text{ is the unit step} \\ &= \frac{N_0}{2}\cdot\begin{cases} u, & u<t \\ t, & u>t \end{cases} \\ R_Y(t,u) &= \frac{N_0}{2}\min(t,u) \end{aligned}$$

Because the correlation function is *not* a function of $t-u$, the process $Y(t)$, which is known as a Wiener process or a random walk process, is non-stationary. When $X(t)$ is Gaussian, the process $Y(t)$ is also known as an *independent increment* process: the so-called increment $Y(t_2)-Y(t_1)$ is statistically independent of the increment $Y(t_4)-Y(t_3)$ so long as the intervals $[t_1,t_2)$, $[t_3,t_4)$ do not overlap. This property follows because of the characteristics of white noise. In fact, all independent increment processes have covariance functions of the form $\sigma^2\min(t,u)$.

2.2.2 The Gaussian Process

A random process $X(t)$ is Gaussian if the joint density of the N amplitudes $X(t_1),\ldots,X(t_N)$ comprise a Gaussian random vector. The elements of the required covariance matrix equal the covariance between the appropriate amplitudes: $K_{ij}=K_X(t_i,t_j)$. Assuming the mean is known, the entire structure of the Gaussian random process is specified once the correlation function or, equivalently, the power spectrum are known. As linear transformations of Gaussian random processes yield another Gaussian process, *linear operations such as differentiation, integration, linear filtering, sampling, and summation with other Gaussian processes result in a Gaussian process*. Thus, many examples of Gaussian processes are generated by passing white Gaussian noise (WGN) through the appropriate linear filter.

2.2.3 Sampling and Random Sequences

The usual Sampling Theorem applies to random processes, with the spectrum of interest being the power spectrum. If stationary process $X(t)$ is bandlimited—$\mathscr{S}_X(f)=0$, $|f|>W$, as long as the sampling interval T satisfies the classic constraint $T<\pi/W$ the sequence $X(lT)$ represents the original process. A sampled process is itself a random process defined over discrete time. Hence, all of the random process notions introduced in the previous section apply to the random sequence $\widetilde{X}(l)\equiv X(lT)$. The correlation functions of these two processes are related as

$$R_{\widetilde{X}}(k)=\mathsf{E}\left[\widetilde{X}(l)\widetilde{X}(l+k)\right]=R_X(kT)\,.$$

We note especially that for distinct samples of a random process to be uncorrelated, the correlation function $R_X(kT)$ must equal zero for all non-zero k. This requirement places severe restrictions on the correlation function (hence the power spectrum) of the original process. One correlation function satisfying this property is derived from the random process which has a bandlimited, constant-valued power spectrum over precisely the frequency region needed to satisfy the sampling criterion. *No other power spectrum satisfying the sampling criterion has this property*. Hence, sampling does not normally yield uncorrelated amplitudes, meaning that *discrete-time white noise* is a rarity. White noise has a correlation function given by $R_{\widetilde{X}}(k)=\sigma^2\delta(k)$, where $\delta(\cdot)$ is the unit sample. The power spectrum of white noise is a constant: $\mathscr{S}_{\widetilde{X}}(f)=\sigma^2$.

2.2.4 The Poisson Process

Some signals have no waveform. Consider the measurement of when lightning strikes occur within some region; the random process is the sequence of event times, which has no intrinsic waveform. Such processes are termed *point processes*, and have been shown [85] to have a simple mathematical structure. Define some quantities first. Let N_t be the number of events that have occurred up to time t (observations are by convention assumed to start at $t = 0$). This quantity is termed the counting process, and has the shape of a staircase function: The counting function consists of a series of plateaus always equal to an integer, with jumps between plateaus occurring when events occur. $N_{t_1,t_2} = N_{t_2} - N_{t_1}$ corresponds to the number of events in the interval $[t_1,t_2)$. Consequently, $N_t = N_{0,t}$. The event times comprise the random vector $\mathbf{W}$; the dimension of this vector is N_t, the number of events that have occurred. The occurrence of events is governed by a quantity known as the *intensity* $\lambda(t;N_t;\mathbf{W})$ of the point process through the probability law

$$\Pr[N_{t,t+\Delta t} = 1 \mid N_t;\mathbf{W}] = \lambda(t;N_t;\mathbf{W})\Delta t + o(\Delta t)$$

for sufficiently small Δt. Here, $o(x)$ represents quantities that approach zero faster than linearly: $\lim_{x\to 0}\frac{o(x)}{x} = 0$. Note that this probability is a conditional probability; it can depend on how many events occurred previously and when they occurred. The intensity can also vary with time to describe non-stationary point processes. The intensity has units of events/s, and it can be viewed as the instantaneous rate at which events occur.

The simplest point process from a structural viewpoint, the Poisson process, has no dependence on process history. A stationary Poisson process results when the intensity equals a constant: $\lambda(t;N_t;\mathbf{W}) = \lambda_0$. Thus, in a Poisson process, a coin is flipped every Δt seconds, with a constant probability of heads (an event) occurring that equals $\lambda_0\Delta t$ and is independent of the occurrence of past (and future) events. When this probability varies with time, the intensity equals $\lambda(t)$, a non-negative signal, and a nonstationary Poisson process results.* From the Poisson process's definition, we can derive the probability laws that govern event occurrence. These fall into two categories: the *count* statistics $\Pr[N_{t_1,t_2} = n]$, the probability of obtaining n events in an interval $[t_1,t_2)$, and the *time of occurrence* statistics $p_{\mathbf{W}^{(n)}}(\mathbf{w})$, the joint distribution of the first n event times in the observation interval. These times form the vector $\mathbf{W}^{(n)}$, the occurrence time vector of dimension n. From these two probability distributions, we can derive the sample function density.

Count statistics. We derive a differentio-difference equation that $\Pr[N_{t_1,t_2} = n]$, $t_1 < t_2$, must satisfy for event occurrence in an interval to be regular and independent of event occurrences in disjoint intervals. Let t_1 be fixed and consider event occurrence in the intervals $[t_1,t_2)$ and $[t_2,t_2+\delta)$, and how these contribute to the occurrence of n events in the union of the two intervals. If k events occur in $[t_1,t_2)$, then $n-k$ must occur in $[t_2,t_2+\delta)$. Furthermore, the scenarios for different values of k are mutually exclusive. Consequently,

$$\begin{aligned}\Pr[N_{t_1,t_2+\delta} = n] &= \sum_{k=0}^{n}\Pr[N_{t_1,t_2} = k, N_{t_2,t_2+\delta} = n-k]\\ &= \Pr[N_{t_2,t_2+\delta} = 0 | N_{t_1,t_2} = n]\Pr[N_{t_1,t_2} = n]\\ &\quad + \Pr[N_{t_2,t_2+\delta} = 1 | N_{t_1,t_2} = n-1]\Pr[N_{t_1,t_2} = n-1]\\ &\quad + \sum_{k=2}^{n}\Pr[N_{t_2,t_2+\delta} = k | N_{t_1,t_2} = n-k]\Pr[N_{t_1,t_2} = n-k]\end{aligned}$$

Because of the independence of event occurrence in disjoint intervals, the conditional probabilities in this expression equal the unconditional ones. When δ is small, only the first two will be significant to first order in δ.

$$\Pr[N_{t_1,t_2+\delta} = n] = \Pr[N_{t_1,t_2} = n]\cdot\big(1-\lambda(t_2)\delta\big) + \Pr[N_{t_1,t_2} = n-1]\cdot\lambda(t_2)\delta$$

Rearranging and taking the obvious limit, we have the equation defining the count statistics.

$$\frac{d\Pr[N_{t_1,t_2} = n]}{dt_2} = -\lambda(t_2)\Pr[N_{t_1,t_2} = n] + \lambda(t_2)\Pr[N_{t_1,t_2} = n-1]$$

*In the literature, stationary Poisson processes are sometimes termed homogeneous, nonstationary ones inhomogeneous.

To solve this equation, we apply a z-transform to both sides. Defining the transform of $\Pr[N_{t_1,t_2} = n]$ to be $P(t_2,z)$,† we have

$$\frac{\partial P(t_2,z)}{\partial t_2} = -\lambda(t_2)(1-z^{-1})P(t_2,z)$$

Applying the boundary condition that $P(t_1,z) = 1$, this simple first-order differential equation has the solution

$$P(t_2,z) = \exp\left\{-(1-z^{-1})\int_{t_1}^{t_2}\lambda(\alpha)\,d\alpha\right\}$$

To evaluate the inverse z-transform, we simply exploit the Taylor series expression for the exponential, and we find that a Poisson probability mass function governs the count statistics for a Poisson process.

$$\boxed{\Pr[N_{t_1,t_2} = n] = \frac{\left(\int_{t_1}^{t_2}\lambda(\alpha)\,d\alpha\right)^n}{n!}\exp\left\{-\int_{t_1}^{t_2}\lambda(\alpha)\,d\alpha\right\}} \tag{2.5}$$

The integral of the intensity occurs frequently, and we succinctly denote it by $\Lambda_{t_1}^{t_2}$.

$$\Lambda_{t_1}^{t_2} \equiv \int_{t_1}^{t_2}\lambda(\alpha)\,d\alpha$$

When the Poisson process is stationary, the intensity equals a constant—$\lambda(t) = \lambda_0$—the count statistics depend only on the difference $t_2 - t_1$: $\Lambda_{t_1}^{t_2} = \lambda_0 \cdot (t_2 - t_1)$.

Time of occurrence statistics. To derive the multivariate distribution of $\mathbf{W}$, we use the count statistics and the independence properties of the Poisson process. The density we seek satisfies

$$\int_{w_1}^{w_1+\delta_1}\cdots\int_{w_n}^{w_n+\delta_n} p_{\mathbf{W}^{(n)}}(\boldsymbol{\upsilon})\,d\boldsymbol{\upsilon} = \Pr\big[W_1 \in [w_1, w_1+\delta_1), \ldots, W_n \in [w_n, w_n+\delta_n)\big]$$

The expression on the right equals the probability that no events occur in $[t_1, w_1)$, one event in $[w_1, w_1+\delta_1)$, no event in $[w_1+\delta_1, w_2)$, etc.. Because of the independence of event occurrence in these disjoint intervals, we can multiply together the probability of these event occurrences, each of which is given by the count statistics.

$$\begin{aligned}
&\Pr\big[W_1 \in [w_1, w_1+\delta_1), \ldots, W_n \in [w_n, w_n+\delta_n)\big] \\
&= e^{-\Lambda_{t_1}^{w_1}} \cdot \Lambda_{w_1}^{w_1+\delta_1} e^{-\Lambda_{w_1}^{w_1+\delta_1}} \cdot e^{-\Lambda_{w_1+\delta_1}^{w_2}} \cdot \Lambda_{w_2}^{w_2+\delta_2} e^{-\Lambda_{w_2}^{w_2+\delta_2}} \cdots \Lambda_{w_n}^{w_n+\delta_n} e^{-\Lambda_{w_n}^{w_n+\delta_n}} \\
&\approx \left(\prod_{k=1}^{n}\lambda(w_k)\delta_k\right) e^{-\Lambda_{t_1}^{w_n}} \quad \text{for small } \delta_k
\end{aligned}$$

From this approximation, we find that the joint distribution of the first n event times equals

$$\boxed{p_{\mathbf{W}^{(n)}}(\mathbf{w}) = \begin{cases} \left(\prod_{k=1}^{n}\lambda(w_k)\right)\exp\left\{-\int_{t_1}^{w_n}\lambda(\alpha)\,d\alpha\right\}, & t_1 \le w_1 \le w_2 \le \cdots \le w_n \\ 0, & \text{otherwise} \end{cases}}$$

Sample function density. For Poisson processes, the sample function density describes the joint distribution of counts and event times within a specified time interval. Thus, it can be written as

$$p_{N_{t_1,t_2},\mathbf{W}}(n;\mathbf{w}) = \Pr[N_{t_1,t_2} = n | W_1 = w_1, \ldots, W_n = w_n]\, p_{\mathbf{W}^{(n)}}(\mathbf{w})$$

The second term in the product equals the distribution derived previously for the time of occurrence statistics. The conditional probability equals the probability that no events occur between w_n and t_2; from the Poisson process's count statistics, this probability equals $\exp\{-\Lambda_{w_n}^{t_2}\}$. Consequently, the sample function density for the Poisson process, be it stationary or not, equals

$$\boxed{p_{N_{t_1,t_2},\mathbf{W}}(n;\mathbf{w}) = \left(\prod_{k=1}^{n}\lambda(w_k)\right)\exp\left\{-\int_{t_1}^{t_2}\lambda(\alpha)\,d\alpha\right\}} \tag{2.6}$$

†Remember, t_1 is fixed and can be suppressed notationally.

Properties. From the probability distributions derived on the previous pages, we can discern many structural properties of the Poisson process. These properties set the stage for delineating other point processes from the Poisson. They, as described subsequently, have much more structure and are much more difficult to handle analytically.

The counting process N_t is an independent increment process. For a Poisson process, the number of events in disjoint intervals are statistically independent of each other, meaning that we have an independent increment process {14}. When the Poisson process is stationary, increments taken over equi-duration intervals are identically distributed as well as being statistically independent. Two important results obtain from this property. First, the counting process's covariance function $K_N(t,u)$ equals $\lambda_0 \min(t,u)$. This close relation to the Wiener waveform process indicates the fundamental nature of the Poisson process in the world of point processes. Note, however, that the Poisson counting process is *not* continuous almost surely. Second, the sequence of counts forms an ergodic process, meaning we can estimate the intensity parameter from observations.

The mean and variance of the number of events in an interval can be easily calculated from the Poisson distribution. Alternatively, we can calculate the characteristic function and evaluate its derivatives. The characteristic function of an increment equals

$$\Phi_{N_{t_1,t_2}}(j\nu) = \exp\left\{\left(e^{j\nu}-1\right)\Lambda_{t_1}^{t_2}\right\}$$

The first two moments and variance of an increment of the Poisson process, be it stationary or not, equal

$$\boxed{\begin{aligned} \mathsf{E}[N_{t_1,t_2}] &= \Lambda_{t_1}^{t_2} \\ \mathsf{E}[N_{t_1,t_2}^2] &= \Lambda_{t_1}^{t_2} + \left(\Lambda_{t_1}^{t_2}\right)^2 \\ \mathsf{var}[N_{t_1,t_2}] &= \Lambda_{t_1}^{t_2} \end{aligned}}$$

Note that the mean equals the variance here, a trademark of the Poisson process.

Poisson process event times form a Markov process. Consider the conditional density $p_{W_n|W_{n-1},\ldots,W_1}(w_n|w_{n-1},\ldots,w_1)$. This density equals the ratio of the event time densities for the n- and $(n-1)$-dimensional event time vectors. Simple substitution yields

$$p_{W_n|W_{n-1},\ldots,W_1}(w_n|w_{n-1},\ldots,w_1) = \lambda(w_n)\exp\left\{-\int_{w_{n-1}}^{w_n}\lambda(\alpha)\,d\alpha\right\}, w_n \geq w_{n-1}$$

Thus, the n^{th} event time depends only on when the $(n-1)^{th}$ event occurs, meaning that we have a Markov process. Note that event times are ordered: The n^{th} event must occur after the $(n-1)^{th}$, etc. When we have a nonstationary Poisson process, the event times form a *nonstationary* Markovian sequence. When the process is stationary, the evolutionary density is exponential. It is this special form of the interevent density that defines a Poisson process.

Inter-event intervals in a Poisson process form a white sequence. Exploiting the previous property, the duration of the n^{th} interval $\tau_n = w_n - w_{n-1}$ does not depend on the lengths of previous (or future) intervals. Consequently, the sequence of inter-event intervals forms a "white" sequence. The sequence may not be identically distributed unless the process is stationary. In the stationary case, inter-event intervals are truly white—they form an IID sequence—and have an exponential distribution.

$$p_{\tau_n}(\tau) = \lambda_0 e^{-\lambda_0 \tau}, \tau \geq 0$$

To show that the exponential density for a white sequence corresponds to the most "random" distribution, Parzen [79] proved that the *ordered* times of n events sprinkled independently and uniformly over a given interval form a stationary Poisson process. If the density of event sprinkling is not uniform, the resulting ordered times constitute a nonstationary Poisson process with an intensity proportional to the sprinkling density.

Doubly stochastic Poisson processes. Here, the intensity $\lambda(t)$ equals a sample function drawn from some waveform process. In waveform processes, the analogous concept does not have nearly the impact it does here. Because intensity waveforms must be non-negative, the intensity process *must* be nonzero mean and non-Gaussian. Assume throughout that the intensity process is stationary for simplicity. This model arises in those situations in which the event occurrence rate clearly varies unpredictably with time. Such processes have the property that the variance-to-mean ratio of the number of events in any interval exceeds one. In the process of deriving this last property, we illustrate the typical way of analyzing doubly stochastic processes: Condition on the intensity equaling a particular sample function, use the statistical characteristics of nonstationary Poisson processes, then "average" with respect to the intensity process. To calculate the expected number N_{t_1,t_2} of events in a interval, we use conditional expected values:

$$\begin{aligned}\mathsf{E}[N_{t_1,t_2}] &= \mathsf{E}\left[\mathsf{E}[N_{t_1,t_2}|\lambda(t), t_1 \le t < t_2]\right] \\ &= \mathsf{E}\left[\int_{t_1}^{t_2} \lambda(\alpha)\, d\alpha\right] \\ &= (t_2 - t_1)\cdot \mathsf{E}[\lambda(t)]\end{aligned}$$

This result can also be written as the expected value of the integrated intensity: $\mathsf{E}[N_{t_1,t_2}] = \mathsf{E}[\Lambda_{t_1}^{t_2}]$. Similar calculations yield the increment's second moment and variance.

$$\begin{aligned}\mathsf{E}[(N_{t_1,t_2})^2] &= \mathsf{E}[\Lambda_{t_1}^{t_2}] + \mathsf{E}[\left(\Lambda_{t_1}^{t_2}\right)^2] \\ \mathsf{var}[N_{t_1,t_2}] &= \mathsf{E}[\Lambda_{t_1}^{t_2}] + \mathsf{var}[\Lambda_{t_1}^{t_2}]\end{aligned}$$

Using the last result, we find that the variance-to-mean ratio in a doubly stochastic process always exceeds unity, equaling one plus the variance-to-mean ratio of the intensity process.

The approach of sample-function conditioning can also be used to derive the density of the number of events occurring in an interval for a doubly stochastic Poisson process. Conditioned on the occurrence of a sample function, the probability of n events occurring in the interval $[t_1, t_2)$ equals (Eq. 2.5, {16})

$$\Pr[N_{t_1,t_2} = n|\lambda(t), t_1 \le t < t_2] = \frac{\left(\Lambda_{t_1}^{t_2}\right)^n}{n!} \exp\left\{-\Lambda_{t_1}^{t_2}\right\}$$

Because $\Lambda_{t_1}^{t_2}$ is a random variable, the unconditional distribution equals this conditional probability averaged with respect to this random variable's density. This average is known as the Poisson Transform of the random variable's density.

$$\boxed{\Pr[N_{t_1,t_2} = n] = \int_0^\infty \frac{\alpha^n}{n!} e^{-\alpha} p_{\Lambda_{t_1}^{t_2}}(\alpha)\, d\alpha}$$

Note that because the density required in this formula is the probability distribution of the *integral* of the intensity. This density function can be difficult to find. In particular, the intensity cannot be a Gaussian process: the intensity must be non-negative for all time. Consequently, we need to find the probability distribution of the integral of some non-Gaussian process.

2.3 Linear Vector Spaces

One of the more powerful tools in statistical communication theory is the abstract concept of a linear vector space. The key result that concerns us is the *representation theorem*: a deterministic time function can be uniquely represented by a sequence of numbers. The stochastic version of this theorem states that a process can be represented by a sequence of uncorrelated random variables. These results will allow us to exploit the theory of hypothesis testing to derive the *optimum* detection strategy.

2.3.1 Basics

Definition *A* linear vector space $\mathscr{S}$ *is a collection of elements called* vectors *having the following properties:*

1. *The vector-addition operation can be defined so that if* $x, y, z \in \mathscr{S}$*:*

(a) $x+y \in \mathscr{S}$ *(the space is closed under addition)*

(b) $x+y=y+x$ *(Commutivity)*

(c) $(x+y)+z=x+(y+z)$ *(Associativity)*

(d) *The zero vector exists and is always an element of* $\mathscr{S}$*. The zero vector is defined by* $x+0=x$*.*

(e) *For each* $x \in \mathscr{S}$*, a unique vector* $(-x)$ *is also an element of* $\mathscr{S}$ *so that* $x+(-x)=0$*, the zero vector.*

2. *Associated with the set of vectors is a set of scalars which constitute an algebraic field. A field is a set of elements which obey the well-known laws of associativity and commutivity for both addition and multiplication. If* a,b *are scalars, the elements* x,y *of a linear vector space have the properties that:*

   (a) $a \cdot x$ *(multiplication by scalar* a*) is defined and* $a \cdot x \in \mathscr{S}$*.*

   (b) $a \cdot (b \cdot x) = (ab) \cdot x$*.*

   (c) *If "1" and "0" denotes the multiplicative and additive identity elements respectively of the field of scalars; then* $1 \cdot x = x$ *and* $0 \cdot x = 0$

   (d) $a(x+y) = ax+ay$ *and* $(a+b)x = ax+bx$*.*

There are many examples of linear vector spaces. A familiar example is the set of column vectors of length N. In this case, we define the sum of two vectors to be:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_1+y_1 \\ x_2+y_2 \\ \vdots \\ x_N+y_N \end{bmatrix}$$

and scalar multiplication to be $a \cdot \text{col}[x_1\, x_2 \cdots x_N] = \text{col}[ax_1\, ax_2 \cdots ax_N]$. All of the properties listed above are satisfied.

A more interesting (and useful) example is the collection of square integrable functions. A square-integrable function $x(t)$ satisfies:

$$\int_{T_i}^{T_f} |x(t)|^2 dt < \infty.$$

One can verify that this collection constitutes a linear vector space. In fact, this space is so important that it has a special name—$L^2(T_i,T_f)$ (read this as *el-two*); the arguments denote the range of integration.

Definition *Let* $\mathscr{S}$ *be a linear vector space. A* subspace *of* $\mathscr{S}$ *is a subset of* $\mathscr{S}$ *which is closed. In other words, if* $x,y \in \mathscr{T}$*, then* $x,y \in \mathscr{S}$ *and all elements of* $\mathscr{T}$ *are elements of* $\mathscr{S}$*, but some elements of* $\mathscr{S}$ *are not elements of* $\mathscr{T}$*. Furthermore, the linear combination* $ax+by \in \mathscr{T}$ *for all scalars* a,b*. A subspace is sometimes referred to as a* closed linear manifold*.*

2.3.2 Inner Product Spaces

A structure needs to be defined for linear vector spaces so that definitions for the length of a vector and for the distance between any two vectors can be obtained. The notions of length and distance are closely related to the concept of an inner product.

Definition *An* inner product *of two real vectors* $x,y \in \mathscr{S}$*, is denoted by* $\langle x,y \rangle$ *and is a* scalar *assigned to the vectors* x *and* y *which satisfies the following properties:*

1. $\langle x,y \rangle = \langle y,x \rangle$
2. $\langle ax,y \rangle = a\langle x,y \rangle$*,* a *is a scalar*
3. $\langle x+y,z \rangle = \langle x,z \rangle + \langle y,z \rangle$*,* z *a vector.*

4. $\langle x,x\rangle > 0$ *unless* $x = 0$. *In this case,* $\langle x,x\rangle = 0$.

Definition *An* inner product space *is a linear vector space in which an inner product can be defined for all elements of the space.*
As an example, an inner product for the space consisting of column matrices can be defined as

$$\langle x,y\rangle = \mathbf{x}^t\mathbf{y} = \sum_{i=1}^{N} x_i y_i\,.$$

The reader should verify that this is indeed a valid inner product (*i.e.*, it satisfies all of the properties given above). It should be noted that this definition of an inner product is not unique: there are other inner product definitions which also satisfy all of these properties. For example, another valid inner product is

$$\langle x,y\rangle_{\mathbf{K}} = \mathbf{x}^t\mathbf{K}\mathbf{y}\,.$$

where $\mathbf{K}$ is an $N\times N$ positive-definite matrix. Choices of the matrix $\mathbf{K}$ which are not positive definite do not yield valid inner products (property 4 is not satisfied). The matrix $\mathbf{K}$ is termed the *kernel* of the inner product. When $\mathbf{K} = \mathbf{I}$, we simplify the notation to $\langle x,y\rangle$.

For the vector space $L^2(T_i,T_f)$, valid inner products are

$$\langle x,y\rangle = \int_{T_i}^{T_f} x(t)y(t)\,dt \quad \langle x,y\rangle_K = \int_{T_i}^{T_f}\int_{T_i}^{T_f} x(t)K(t,u)y(u)\,dt\,du$$

where $K(t,u)$ is a positive-definite kernel function: $\langle x,x\rangle_K > 0$ for all $x \neq 0$. The identity kernel in this case is $\delta(t-u)$.

Definition *The* norm *of a vector* x*, an element of the inner product space* $\mathscr{S}$*, is denoted by* $\|x\|$ *and equals*

$$\|x\| = \langle x,x\rangle^{1/2} \tag{2.7}$$

Because of property 4 of an inner product, the norm of a vector is always greater than zero unless the vector is identically zero. The norm of a vector is related to the notion of the *length* of a vector. For example, if the vector x is multiplied by a constant scalar a, the norm of the vector is also multiplied by $|a|$.

$$\|ax\| = \langle ax,ax\rangle^{1/2} = |a|\|x\|$$

In other words, "longer" vectors ($|a| > 1$) have larger norms. For the space $\mathscr{S}$ consisting of column matrices, the identity-kernel norm of a vector is given by

$$\|x\| = \left(\sum_{i=1}^{N} x_i^2\right)^{1/2}.$$

This choice corresponds to the Cartesian definition of the length of a vector. A norm can also be defined when the inner product contains a kernel. In this case, the norm is written $\|x\|_{\mathbf{K}}$ for clarity: $\|x\|_K = \langle x,x\rangle_K^{1/2}$.

One of the fundamental properties of inner product spaces is the *Schwarz inequality.*

$$|\langle x,y\rangle| \leq \|x\|\,\|y\| \tag{2.8}$$

This is one of the most important inequalities we shall encounter. To demonstrate this inequality, consider the norm squared of $x+ay$.

$$\|x+ay\|^2 = \langle x+ay,x+ay\rangle = \|x\|^2 + 2a\langle x,y\rangle + a^2\|y\|^2$$

Let $a = -\langle x,y\rangle/\|y\|^2$. In this case:

$$\begin{aligned}\|x+ay\|^2 &= \|x\|^2 - 2\frac{|\langle x,y\rangle|^2}{\|y\|^2} + \frac{|\langle x,y\rangle|^2}{\|y\|^4}\|y\|^2\\ &= \|x\|^2 - \frac{|\langle x,y\rangle|^2}{\|y\|^2}\end{aligned}$$

As the left hand side of this result is non-negative, the right-hand side is lower-bounded by zero. The Schwarz inequality of Eq. 2.8 is thus obtained. Note that equality occurs *only* when $x = -ay$: the vectors x, y are proportional to each other. Because the Schwarz inequality was proved only using the properties of inner products, it applies to both column matrices and finite-energy signals and when a kernel is present.

Definition *Two vectors are said to be* orthogonal *if the inner product of the vectors is zero:* $\langle x,y\rangle = 0$.

Consistent with these results is the concept of the "angle" between two vectors. The cosine of this angle is defined by:

$$\cos(x,y) = \frac{\langle x,y\rangle}{\|x\|\,\|y\|}$$

Because of the Schwarz inequality, $|\cos(x,y)| \leq 1$. The angle between orthogonal vectors is $\pm\pi/2$ and the angle between vectors satisfying Eq. 2.8 with equality ($x \propto y$) is zero (the vectors are parallel to each other).

2.3.3 Normed Spaces

A norm can be defined without being based on an inner product, provided it forms a valid distance measure.

Definition *The distance d between two vectors is taken to be the norm of the difference of the vectors.*

$$d(x,y) = \|x-y\|$$

This distance measure (or metric) must have the following properties:

1. $d(x,y) = d(y,x)$ *(Distance does not depend on how it is measured.)*
2. $d(x,y) = 0 \implies x = y$ *(Zero distance means equality.)*
3. $d(x,y) \leq d(x,z) + d(z,y)$ *or* $\|x+y\| \leq \|x\| + \|y\|$ *(Triangle inequality)*

These properties follow from those for an inner product but we don't need an inner product to define a norm. For example, the so-called p-norm ($p \geq 1$) is defined for column vectors to be

$$\|x\|_p = \left(\sum_{n=1}^{N} |x_n|^p\right)^{1/p}$$

This definition is easily extended to functions: $\|f\|_p = \left(\int |f(x)|^p\,dx\right)^{1/p}$. If $0 < p < 1$ we no longer have a norm.* We define the normed space $L^p(T_i,T_f)$ as having a p-norm. For example, $L^1(T_i,T_f)$ has a norm for column matrices given by $\|x\| = \sum_n |x_n|$ for which no inner product can be found. *If* an inner product underlies the norm ($\|x\|^2 = \langle x,x\rangle$), it can be found from the norm: $\|x+y\|^2 - \|x-y\|^2 = 4\langle x,y\rangle$. The p-norm of a vector is a non-increasing function of p: $\|x\|_{p+q} \leq \|x\|_p$, $q \geq 0$. Consequently, $\|x\|_2 \leq \|x\|_1$. To see this result, trace out the "unit sphere" $\|x\|_p = 1$ for $p = 1$ and $p = 2$.

We use a distance measure to define what we mean by convergence. When we say the sequence of vectors $\{x_n\}$ converges to x ($x_n \to x$), we mean

$$\lim_{n\to\infty} \|x_n - x\| = 0$$

Normed, linear vector spaces provide a formal framework for discussing error. We can define ε_n to be the error between x and the sequence x_n: $\varepsilon_n = x_n - x$. The norm of the error $\|\varepsilon_n\|$ captures the strength of the error for each n. When we consider estimation error in sequel, we will frequently use the L^2 norm as our measure of the error's strength.

*When $0 < p < 1$, the triangle inequality is violated.

2.4 Distance between Probability Densities

How different are two probability densities? Is a Laplacian density "closer" to a Gaussian than the hyperbolic secant density?† *If* we could define a linear vector space for probability densities, we could simply calculate the distance between any two densities using the norm of their difference. Unfortunately, probability densities cannot form a linear vector space: the properties of a probability density prevent this mathematical structure. As described in §2.3, the sum of two vectors must be a vector; the sum of two densities is clearly not a density since the integral of the sum does not equal one. Also, the "zero vector" must also be a vector; again, the integral of a density must be one. Consequently, we cannot use linear vector space theory to characterize how different two densities are.

The discrepancy away from a nominal density is difficult to define. Any discrepancy expression must be such that the result cannot be negative and have unity area. One workable solution is to use a *mixture model*.

$$p(x) = (1-a)p_0(x) + ap_d(x),\ 0 \le a \le 1$$

Here $p_0(x)$ is the nominal density and $p_d(x)$ is the discrepancy (itself a density) between $p(x)$ and the nominal. Here, the parameter a would be a measure of the discrepancy.*

However, it is certainly not true that an arbitrary density can be described by a mixture model from any nominal. For example, a uniform density cannot be described as a mixture of a discrepancy density with a Gaussian. We will need a way to characterize how far apart two densities are — what is the distance between them — in a meaningful way. Rather than a linear vector space, the alternative formalism for describing densities is to create a *manifold*. In simple terms, a manifold is a surface, each point on which has some particular property (like being a density). One can create a manifold containing *all* densities. Typically, manifolds are not flat; they are "bumpy." The distance between any two points on the manifold is taken as the length of the shortest path that joins the two points. It can be shown that the shortest path has a "minimum work" property: the shortest path, known as the *geodesic*, tends to follow the valleys. For the manifold of probability densities, the geodesic joining p_0 and p_1 is given by p_s, defined to be $p_0^{1-s}p_1^s/J(s)$, where $J(s)$ provides the normalization so that p_s integrates to one. The parameter s ranges over $[0,1]$, defining a path that passes from p_0 when $s=0$ to p_1 when $s=1$.

Among the distance choices for the manifold of probability densities that have proved fruitful are the *Kullback-Leibler* and *Chernoff* distances. Respectively,

$$\mathscr{D}(p_1\|p_0) = \int p_1(x)\log\frac{p_1(x)}{p_0(x)}\,dx \tag{2.9}$$

$$\mathscr{C}(p_0,p_1) = -\min_{0\le s\le 1}\log J(s),\ J(s) = \int [p_0(x)]^{1-s}\,[p_1(x)]^s\,dx \tag{2.10}$$

The Kullback-Leibler distance is a special case of an Ali-Silvey distance (Appendix C {247}).† These distances have the following properties.

1. $\mathscr{D}(p_1\|p_0) \ge 0$ and $\mathscr{C}(p_1,p_0) \ge 0$. Furthermore, these distances equal zero only when $p_0(X) = p_1(X)$. The Kullback-Leibler and Chernoff distances are always non-negative, with zero distance occurring only when the probability distributions are the same. To prove this fact, we focus on showing that the Kullback-Leibler distance is non-negative. It follows from subsequent discussion that the Chernoff distance equals a Kullback-Leibler distance; consequently, the Chernoff distance is non-negative. The derivation relies on Jensen's inequality.

†Appendix A {229} describes the form of these densities.

*It should be noted that mixture models are also used to describe "unusual" behavior of real data. For example, $p_0(x)$ could be a zero-mean Gaussian density and the density $p_d(x)$ another zero-mean Gaussian density having a much larger variance. The parameter a can be interpreted as a probability: the random variable having density $p(x)$ is a Gaussian having a nominal variance with probability $(1-a)$ or a Gaussian with a much larger variance with probability a.

†If it were not for the optimization in the definition of the Chernoff distance, it too would be in the Ali-Silvey class.

Aside: Jensen's Inequality

Jensen's inequality is frequently used in probability theory and in information theory. It states

$$\boxed{\mathsf{E}[f(X)] \geq f(\mathsf{E}[X]),\ f(\cdot) \text{ a convex function}} \tag{2.11}$$

Furthermore, if $f(\cdot)$ is strictly convex, equality applies only if $p_X(x) = \delta(x - x_0)$ so that $X = \mathsf{E}[X]$. If $g(\cdot)$ is concave, the inequality reverses.

$$\boxed{\mathsf{E}[g(X)] \leq g(\mathsf{E}[X]),\ g(\cdot) \text{ a concave function}} \tag{2.12}$$

The proof of this result is easy for discrete probability distributions. First of all, if the random variable can assume only two values,

$$\mathsf{E}[f(X)] = p_1 f(X_1) + p_2 f(X_2),\ p_2 = 1 - p_1$$

The definition of a convex function is that it lies below a straight line connecting the function's values at any two points within the function's range. Consequently,

$$p_1 f(X_1) + p_2 f(X_2) \geq f(p_1 X_1 + p_2 X_2)$$

Assume Jensen's inequality holds for a random variable assuming n values.

$$\sum_{k=1}^{n} p_k f(X_k) \geq f\left(\sum_{k=1}^{n} p_x X_k\right)$$

We use induction to show that the inequality must also hold for $n+1$ values.

$$\begin{aligned}\sum_{k=1}^{n+1} p_k f(X_k) &= \sum_{k=1}^{n} p_k f(X_k) + p_{n+1} f(X_{n+1}) \\ &= (1 - p_{n+1}) \sum_{k=1}^{n} \frac{p_k}{1 - p_{n+1}} f(X_k) + p_{n+1} f(X_{n+1})\end{aligned}$$

Because this expression amounts to the two-point result,

$$\geq f\left((1 - p_{n+1}) \sum_{k=1}^{n} \frac{p_k}{1 - p_{n+1}} X_k + p_{n+1} X_{n+1}\right) = f\left(\sum_{k=1}^{n+1} p_x X_k\right)$$

and Jensen's inequality is proven for the discrete-valued random variable case. The inequality also applies when the random variable has a density. The proof follows naturally from the discrete case result.

With this result, the derivation is easy.

$$\begin{aligned}\mathscr{D}(p_1 \| p_0) &= \int p_1(\alpha) \log \frac{p_1(\alpha)}{p_0(\alpha)}\, d\alpha \\ &= \int p_1(\alpha) \left[-\log \frac{p_0(\alpha)}{p_1(\alpha)}\right] d\alpha \\ &\geq -\log\left(\int p_1(\alpha) \cdot \frac{p_0(\alpha)}{p_1(\alpha)}\, d\alpha\right) \quad [-\log(\cdot) \text{ is convex}] \\ &= -\log\left(\int p_0(\alpha)\, d\alpha\right) = -\log(1) = 0\end{aligned}$$

Furthermore, this quantity equals zero *only* when the two probability distributions equal each other.

2. $\mathscr{D}(p_1\|p_0)=\infty$ whenever, for some x, $p_0(x)=0$ and $p_1(x)\neq 0$. If $p_1(x)=0$, the value of $p_1(x)\log\frac{p_1(x)}{p_0(x)}$ is defined to be zero.
3. When the underlying stochastic quantities are random vectors having statistically independent components with respect to both p_0 and p_1, the Kullback-Leibler distance equals the sum of the component distances. Stated mathematically, if $p_0(\mathbf{x})=\prod_l p_0(X_l)$ and $p_1(\mathbf{x})=\prod_l p_1(x_l)$,
$$\mathscr{D}(p_1(\mathbf{x})\|p_0(\mathbf{x}))=\sum_l \mathscr{D}(p_1(x_l)\|p_0(x_l))\ . \tag{2.13}$$
The Chernoff distance does *not* have this *additivity* property unless the components are identically distributed.
4. $\mathscr{D}(p_1\|p_0)\neq\mathscr{D}(p_0\|p_1)$; $\mathscr{C}(p_0,p_1)=\mathscr{C}(p_1,p_0)$.
The Kullback-Leibler distance is usually not a symmetric quantity. In some special cases, it can be symmetric (like the just described Gaussian example), but symmetry cannot, and should not, be expected. The Chernoff distance is always symmetric.
5. $\mathscr{D}(p(X_1,X_2)\|p(X_1)p(X_2))=I(X_1;X_2)$.
The Kullback-Leibler distance between a joint probability density and the product of the marginal distributions equals what is known in information theory as the *mutual information* between the random variables X_1, X_2. From the properties of the Kullback-Leibler distance, we see that the mutual information equals zero only when the random variables are statistically independent.

These quantities are not actually distances in the strict mathematical sense {21}. The Kullback-Leibler distance $\mathscr{D}(\cdot\|\cdot)$ is not symmetric in its arguments and the Chernoff distance $\mathscr{C}(\cdot,\cdot)$ does not obey the triangle inequality {21}. Nonetheless, the name "distance" is frequently attached to them for reasons that will become clear later. $\mathscr{D}(p_1\|p_0)$ is interpreted as the distance p_1 is from p_0. Because of the asymmetry, the distance between two points on the manifold of probability densities depends on which point is used as the origin.*

The Kullback-Leibler and Chernoff distances are related in an important way. Let $s^*=\arg\min_s J(s)$, the value of s that defines the Chernoff distance in (2.10). s^* also defines the density p_s is equi-distant from p_0 and p_1.
$$\mathscr{D}(p_{s^*}\|p_0)=\mathscr{D}(p_{s^*}\|p_1)$$
To show this result, we first consider how to find the Chernoff distance (Eq. (2.10)). The minimum of $\log J(s)$ occurs at the minimum of $J(s)$. Interestingly, $J(s)$ is strictly convex, which means its minimum can always be found by setting its derivative to zero. The location of the minimum, s^*, is the solution of
$$\left.\frac{dJ(s)}{ds}\right|_{s=s^*}=\int p_0(x)\left[\frac{p_1(x)}{p_0(x)}\right]^{s^*}\log\frac{p_1(x)}{p_0(x)}\,dx=0 \tag{*}$$
No closed form solution exists in general. The equi-distant point from the viewpoint of the Kullback-Leibler distance must satisfy
$$\underbrace{\int\frac{p_0(x)\left[\frac{p_1(x)}{p_0(x)}\right]^s}{J(s)}\log\left[\frac{p_1(x)}{p_0(x)}\right]^s dx-\log J(s)}_{\mathscr{D}(p_s\|p_0)}=\underbrace{\int\frac{p_0(x)\left[\frac{p_1(x)}{p_0(x)}\right]^s}{J(s)}\log\left[\frac{p_1(x)}{p_0(x)}\right]^{s-1} dx-\log J(s)}_{\mathscr{D}(p_s\|p_1)} \tag{**}$$
Combining (**) into a single integral yields the equation (*) that defines s^*. In addition, equation (*) means that the integral on the left side of equation (**) equals zero when $s=s^*$, which means that $\mathscr{D}(p_{s^*}\|p_0)=-\log J(s^*)$. Consequently, the Chernoff distance equals the equi-distant Kullback-Leibler distance.†
$$\mathscr{C}(p_0,p_1)=\mathscr{D}(p_{s^*}\|p_0)=\mathscr{D}(p_{s^*}\|p_1)$$
We see that the Chernoff distance between two densities is upper-bounded by the Kullback-Leibler distance between two densities regardless of which density is chosen as the reference.

*For Riemannian manifolds, distances are always symmetric, as they should be. For technical reasons, the manifold of probability densities is non-Riemannian.

†It is this result that implies that the Chernoff distance is non-negative.

2.5 Hilbert Spaces

Definition *A Hilbert space $\mathscr{H}$ is an inner product linear vector space which contains all of its limit points: if $\{x_n\}$ is any sequence of elements in $\mathscr{H}$ that converges to x ($\lim_{n\to\infty}\|x_n - x\| = 0$), then x is also contained in $\mathscr{H}$. x is termed the* limit point *of the sequence.*

Example

Let the space consist of all rational numbers. Let the inner product be simple multiplication: $\langle x,y\rangle = xy$. However, the limit point of the sequence $x_n = 1+1+1/2!+\cdots+1/n!$ is not a rational number. Consequently, this space is *not* a Hilbert space. However, if we define the space to consist of all finite numbers, we have a Hilbert space.

Definition *If $\mathscr{Y}$ is a subspace of $\mathscr{H}$, the vector x is* orthogonal *to the subspace $\mathscr{Y}$ for every $y \in \mathscr{Y}$, $\langle x,y\rangle = 0$.*
We now arrive at a fundamental theorem.

Theorem *Let $\mathscr{H}$ be a Hilbert space and $\mathscr{Y}$ a subspace of it. Any element $x \in H$ has the unique decomposition $x = y+z$, where $y \in \mathscr{Y}$ and z is orthogonal to $\mathscr{Y}$. Furthermore, $\|x-y\| = \min_{v\in\mathscr{Y}}\|x-v\|$: the distance between x and all elements of $\mathscr{Y}$ is minimized by the vector y. This element y is termed the* projection *of x onto $\mathscr{Y}$.*

Geometrically, $\mathscr{Y}$ is a line or a plane passing through the origin. Any vector x can be expressed as the linear combination of a vector lying in $\mathscr{Y}$ and a vector orthogonal to y. This theorem is of extreme importance in linear estimation theory and plays a fundamental role in detection theory.

2.5.1 Separable Vector Spaces

Definition *A Hilbert space $\mathscr{H}$ is said to be* separable *if there exists a set of vectors $\{\phi_i\}$, $i = 1,\ldots,$ elements of $\mathscr{H}$, that express* every *element $x \in \mathscr{H}$ as*

$$x = \sum_{i=1}^{\infty} a_i\phi_i\,, \tag{2.14}$$

where a_i are scalar constants associated with ϕ_i and x and where "equality" is taken to mean that the distance between each side becomes zero as more terms are taken in the right.

$$\lim_{m\to\infty}\left\| x - \sum_{i=1}^{m} a_i\phi_i \right\| = 0$$

The set of vectors $\{\phi_i\}$ are said to form a *complete* set if the above relationship is valid and is said to form a *basis* for the space $\mathscr{H}$. The elements of a basis for a space are taken to be linearly independent. *Linear independence* implies that the expression of the zero vector by a basis can only be made by zero coefficients.

$$\sum_{i=1}^{\infty} a_i\phi_i = 0 \Leftrightarrow a_i = 0\,,\ i = 1,\ldots$$

The *representation theorem* states simply that separable vector spaces exist. The representation of the vector x is the sequence of coefficients $\{a_i\}$.

Example

The space consisting of column matrices of length N is easily shown to be separable. Let the vector ϕ_i be given a column matrix having a one in the i^{th} row and zeros in the remaining rows:

$\phi_i = \text{col}[0,\ldots,0,1,0,\ldots,0]$. This set of vectors $\{\phi_i\}$, $i = 1,\ldots,N$ constitutes a basis for the space. Obviously if the vector x is given by $x = \text{col}[x_1\, x_2 \ldots x_N]$, it may be expressed as

$$x = \sum_{i=1}^{N} x_i \phi_i$$

since $a_i = x_i$ using the basis vectors just defined.

In general, the upper limit on the sum in Eq. 2.14 is infinite. For the previous example, the upper limit is finite. The number of basis vectors that is *required* to express every element of a separable space in terms of Eq. 2.14 is said to be the *dimension* of the space. In this example, the dimension of the space is N. There exist separable vector spaces for which the dimension is infinite.

Definition *The basis for a separable vector space is said to be an* orthonormal *basis if the elements of the basis satisfy the following two properties:*

- *The inner product between distinct elements of the basis is zero (i.e., the elements of the basis are mutually orthogonal).*
$$\langle \phi_i, \phi_j \rangle = 0\,,\ i \neq j$$
- *The norm of each element of a basis is one (normality).*
$$\|\phi_i\| = 1\,,\ i = 1,\ldots$$

For example, the basis given above for the space of N-dimensional column matrices is orthonormal. For clarity, two facts must be explicitly stated. First, not every basis is orthonormal. If the vector space is separable, a complete set of vectors can be found; however, this set does not have to be orthonormal to be a basis. Secondly, not every set of orthonormal vectors can constitute a basis. When the vector space L^2 is discussed in detail, this point will be illustrated.

Despite these qualifications, an orthonormal basis exists for every separable vector space. There is an explicit algorithm—the *Gram-Schmidt procedure*—for deriving an orthonormal set of functions from a complete set. Let $\{\phi_i\}$ denote a basis; the orthonormal basis $\{\psi_i\}$ is sought. The Gram-Schmidt procedure is:

1. $\psi_1 = \phi_1 / \|\phi_1\|$.
   This step makes ψ_1 have unit length.
2. $\psi_2' = \phi_2 - \langle \psi_1, \phi_2 \rangle \psi_1$.
   Consequently, the inner product between ψ_2' and ψ_1 is zero. We obtain ψ_2 from ψ_2' forcing the vector to have unit length.

$2'$. $\psi_2 = \psi_2' / \|\psi_2'\|$.
   The algorithm now generalizes.

k. $\psi_k' = \phi_k - \sum_{i=1}^{k-1} (\psi_i, \phi_k) \psi_i$

k'. $\psi_k = \psi_k' / \|\psi_k'\|$

By construction, this new set of vectors is an orthonormal set. As the original set of vectors $\{\phi_i\}$ is a complete set, and, as each ψ_k is just a linear combination of ϕ_i, $i = 1,\ldots,k$, the derived set $\{\psi_i\}$ is also complete. Because of the existence of this algorithm, a basis for a vector space is usually assumed to be orthonormal.

A vector's representation with respect to an orthonormal basis $\{\phi_i\}$ is easily computed. The vector x may be expressed by:

$$x = \sum_{i=1}^{\infty} a_i \phi_i \tag{2.15}$$

$$a_i = \langle x, \phi_i \rangle \tag{2.16}$$

This formula is easily confirmed by substituting Eq. 2.15 into Eq. 2.16 and using the properties of an inner product. Note that the exact element values of a given vector's representation depends upon both the vector *and* the choice of basis. Consequently, a meaningful specification of the representation of a vector must include the definition of the basis.

The mathematical representation of a vector (expressed by equations 2.15 and 2.16) can be expressed geometrically. This expression is a generalization of the Cartesian representation of numbers. Perpendicular axes are drawn; these axes correspond to the orthonormal basis vector used in the representation. A given vector is representation as a point in the "plane" with the value of the component along the ϕ_i axis being a_i.

An important relationship follows from this mathematical representation of vectors. Let x and y be any two vectors in a separable space. These vectors are represented with respect to an orthonormal basis by $\{a_i\}$ and $\{b_i\}$, respectively. The inner product $\langle x,y\rangle_K$ is related to these representations by:

$$\langle x,y\rangle_K = \sum_{i=1}^{\infty} a_i b_i$$

This result is termed *Parseval's Theorem*. Consequently, the inner product between any two vectors can be computed from their representations. A special case of this result corresponds to the Cartesian notion of the length of a vector; when $x = y$, Parseval's relationship becomes:

$$\|x\| = \left[\sum_{i=1}^{\infty} a_i^2\right]^{1/2}$$

These two relationships are key results of the representation theorem. The implication is that any inner product computed from vectors can also be computed from their representations. There are circumstances in which the latter computation is more manageable than the former and, furthermore, of greater theoretical significance.

2.5.2 The Vector Space L^2

Special attention needs to be paid to the vector space $L^2(T_i,T_f)$: the collection of functions $x(t)$ which are square-integrable over the interval (T_i,T_f):

$$\int_{T_i}^{T_f} |x(t)|^2\,dt < \infty$$

An inner product can be defined for this space as:

$$\langle x,y\rangle = \int_{T_i}^{T_f} x(t)y(t)\,dt \tag{2.17}$$

Consistent with this definition, the length of the vector $x(t)$ is given by

$$\|x\| = \left[\int_{T_i}^{T_f} |x(t)|^2\,dt\right]^{1/2}$$

Physically, $\|x\|^2$ can be related to the energy contained in the signal over (T_i,T_f). This space is a Hilbert space. If T_i and T_f are both finite, an orthonormal basis is easily found which spans it. For simplicity of notation, let $T_i = 0$ and $T_f = T$. The set of functions defined by:

$$\begin{aligned}\phi_{2i-1}(t) &= \left(\frac{2}{T}\right)^{1/2} \cos\frac{2\pi(i-1)t}{T}\\ \phi_{2i}(t) &= \left(\frac{2}{T}\right)^{1/2} \sin\frac{2\pi it}{T}\end{aligned} \tag{2.18}$$

is complete over the interval $(0,T)$ and therefore constitutes a basis for $L^2(0,T)$. By demonstrating a basis, we conclude that $L^2(0,T)$ is a separable vector space. The representation of functions with respect to this

basis corresponds to the well-known Fourier series expansion of a function. As most functions require an infinite number of terms in their Fourier series representation, this space is infinite dimensional.

There also exist orthonormal sets of functions that do *not* constitute a basis. For example, the set $\{\phi_i(t)\}$ defined by:

$$\phi_i(t) = \begin{cases} \frac{1}{T} & iT \le t < (i+1)T \\ 0 & \text{otherwise} \end{cases} \qquad i = 0, 1, \ldots$$

over $L^2(0,\infty)$. The members of this set are normal (unit norm) and are mutually orthogonal (no member overlaps with any other). Consequently, this set is an orthonormal set. However, it does not constitute a basis for $L^2(0,\infty)$. Functions piecewise constant over intervals of length T are the only members of $L^2(0,\infty)$ which can be represented by this set. Other functions such as $e^{-t}\mathrm{u}(t)$ cannot be represented by the $\{\phi_i(t)\}$ defined above. Consequently, *orthonormality of a set of functions does not guarantee completeness*.

While $L^2(0,T)$ is a separable space, examples can be given in which the representation of a vector in this space is not precisely equal to the vector. More precisely, let $x(t) \in L^2(0,T)$ and the set $\{\phi_i(t)\}$ be defined by Eq. (2.18). The fact that $\{\phi_i(t)\}$ formss a basis for the space implies:

$$\left\| x(t) - \sum_{i=1}^{\infty} a_i \phi_i(t) \right\| = 0 \quad \text{where } a_i = \int_0^T x(t)\phi_i(t)\,dt\,.$$

In particular, let $x(t)$ be:

$$x(t) = \begin{cases} 1 & 0 \le t \le T/2 \\ 0 & T/2 < t < T \end{cases}$$

Obviously, this function is an element of $L^2(0,T)$. However, the representation of this function is not equal to 1 at $t = T/2$. In fact, the peak error never decreases as more terms are taken in the representation. In the special case of the Fourier series, the existence of this "error" is termed the *Gibbs phenomenon*. However, this "error" has zero norm in $L^2(0,T)$; consequently, the Fourier series expansion of this function is equal to the function in the sense that the function and its expansion have zero distance between them. However, one of the axioms of a valid inner product is that if $\|e\| = 0 \implies e = 0$. The condition is satisfied, but the conclusion does not seem to be valid. Apparently, valid elements of $L^2(0,T)$ can be defined which are nonzero but have zero norm. An example is

$$e = \begin{cases} 1 & t = T/2 \\ 0 & \text{otherwise} \end{cases}$$

So as not to destroy the theory, the most common method of resolving the conflict is to weaken the definition of equality. The essence of the problem is that while two vectors x and y can differ from each other and be zero distance apart, the difference between them is "trivial". This difference has zero norm which, in L^2, implies that the magnitude of $(x-y)$ integrates to zero. Consequently, the vectors are essentially equal. This notion of equality is usually written as $x = y$ a.e. (x equals y *almost everywhere*). With this convention, we have:

$$\|e\| = 0 \implies e = 0 \text{ a.e.}$$

Consequently, the error between a vector and its representation is zero almost everywhere.

Weakening the notion of equality in this fashion might seem to compromise the utility of the theory. However, if one suspects that two vectors in an inner product space are equal (*e.g.*, a vector and its representation), it is quite difficult to prove that they are strictly equal (and as has been seen, this conclusion may not be valid). Usually, proving they are equal almost everywhere is much easier. While this weaker notion of equality does not imply strict equality, one can be assured that any difference between them is insignificant. The measure of "significance" for a vector space is expressed by the definition of the norm for the space.

2.5.3 A Hilbert Space for Stochastic Processes

The result of primary concern here is the construction of a Hilbert space for stochastic processes. The space consisting of random variables X having a finite mean-square value is (almost) a Hilbert space with inner

product $\mathsf{E}[XY]$. Consequently, the distance between two random variables X and Y is

$$d(X,Y) = \left\{\mathsf{E}[(X-Y)^2]\right\}^{1/2}$$

Now $d(X,Y) = 0 \implies \mathsf{E}[(X-Y)^2] = 0$. However, this does not imply that $X = Y$. Those sets with probability zero appear again. Consequently, we do not have a Hilbert space unless we agree $X = Y$ means $\Pr[X = Y] = 1$.

Let $X(t)$ be a process with $\mathsf{E}[X^2(t)] < \infty$. For each t, $X(t)$ is an element of the Hilbert space just defined. Parametrically, $X(t)$ is therefore regarded as a "curve" in a Hilbert space. This curve is continuous if

$$\lim_{t\to u} \mathsf{E}[\big(X(t) - X(u)\big)^2] = 0$$

Processes satisfying this condition are said to be *continuous in the quadratic mean*. The vector space of greatest importance is analogous to $L^2(T_i,T_f)$ previously defined. Consider the collection of real-valued stochastic processes $X(t)$ for which

$$\int_{T_i}^{T_f} \mathsf{E}[X(t)^2]\,dt < \infty$$

Stochastic processes in this collection are easily verified to constitute a linear vector space. Define an inner product for this space as:

$$\mathsf{E}[\langle X(t), Y(t)\rangle] = \mathsf{E}\left[\int_{T_i}^{T_f} X(t)Y(t)\,dt\right]$$

While this equation is a valid inner product, the left-hand side will be used to denote the inner product instead of the notation previously defined. We take $\langle X(t), Y(t)\rangle$ to be the *time-domain inner product* as in Eq. (2.17). In this way, the deterministic portion of the inner product and the expected value portion are explicitly indicated. This convention allows certain theoretical manipulations to be performed more easily.

One of the more interesting results of the theory of stochastic processes is that the normed vector space for processes previously defined is separable. Consequently, there exists a complete (and, by assumption, orthonormal) set $\{\phi_i(t)\}, i = 1,\ldots$ of deterministic (nonrandom) functions which constitutes a basis. A process in the space of stochastic processes can be represented as

$$X(t) = \sum_{i=1}^{\infty} X_i \phi_i(t)\,, \quad T_i \le t \le T_f\,,$$

where $\{X_i\}$, the representation of $X(t)$, is a sequence of random variables given by

$$X_i = \langle X(t), \phi_i(t)\rangle \quad \text{or} \quad X_i = \int_{T_i}^{T_f} X(t)\phi_i(t)\,dt\,.$$

Strict equality between a process and its representation cannot be assured. Not only does the analogous issue in $L^2(0,T)$ occur with respect to representing individual sample functions, but also sample functions assigned a zero probability of occurrence can be troublesome. In fact, the ensemble of any stochastic process can be augmented by a set of sample functions that are not well-behaved (e.g., a sequence of impulses) but have probability zero. In a practical sense, this augmentation is trivial: such members of the process cannot occur. Therefore, one says that two processes $X(t)$ and $Y(t)$ are equal almost everywhere if the distance between $\|X(t) - Y(t)\|$ is zero. The implication is that any lack of strict equality between the processes (strict equality means the processes match on a sample-function-by-sample-function basis) is "trivial".

2.5.4 Karhunen-Loève Expansion

The representation of the process, $X(t)$, is the sequence of random variables X_i. The choice basis of $\{\phi_i(t)\}$ is unrestricted. Of particular interest is to restrict the basis functions to those which make the $\{X_i\}$ *uncorrelated* random variables. When this requirement is satisfied, the resulting representation of $X(t)$ is termed the *Karhunen-Loève* expansion. Mathematically, we require $\mathsf{E}[X_iX_j] = \mathsf{E}[X_i]\,\mathsf{E}[X_j]$, $i \neq j$. This requirement can

be expressed in terms of the correlation function of $X(t)$.

$$\begin{aligned}\mathsf{E}[X_iX_j] &= \mathsf{E}\left[\int_0^T X(\alpha)\phi_i(\alpha)\,d\alpha \int_0^T X(\beta)\phi_j(\beta)\,d\beta\right] \\ &= \int_0^T\int_0^T \phi_i(\alpha)\phi_j(\beta)R_X(\alpha,\beta)\,d\alpha\,d\beta\end{aligned}$$

As $\mathsf{E}[X_i]$ is given by

$$\mathsf{E}[X_i] = \int_0^T m_X(\alpha)\phi_i(\alpha)\,d\alpha\,,$$

our requirement becomes

$$\int_0^T\int_0^T \phi_i(\alpha)\phi_j(\beta)R_X(\alpha,\beta)\,d\alpha\,d\beta = \int_0^T m_X(\alpha)\phi_i(\alpha)\,d\alpha \int_0^T m_X(\beta)\phi_j(\beta)\,d\beta\,,\ i\neq j\,.$$

Simple manipulations result in the expression

$$\int_0^T \phi_i(\alpha)\left[\int_0^T K_X(\alpha,\beta)\phi_j(\beta)\,d\beta\right]d\alpha = 0\,,\ i\neq j\,.$$

When $i=j$, the quantity $\mathsf{E}[X_i^2]-\mathsf{E}^2[X_i]$ is just the variance of X_i. Our requirement is obtained by satisfying

$$\int_0^T \phi_i(\alpha)\left[\int_0^T K_X(\alpha,\beta)\phi_j(\beta)\,d\beta\right]d\alpha = \lambda_i\delta_{ij}$$

or

$$\int_0^T \phi_i(\alpha)g_j(\alpha)\,d\alpha = 0\,,\ i\neq j\,,$$

where

$$g_j(\alpha) = \int_0^T K_X(\alpha,\beta)\phi_j(\beta)\,d\beta\,.$$

Furthermore, this requirement must hold for each j which differs from the choice of i. A choice of a function $g_j(\alpha)$ satisfying this requirement is a function which is proportional to $\phi_j(\alpha)$: $g_j(\alpha)=\lambda_j\phi_j(\alpha)$. Therefore,

$$\boxed{\int_0^T K_X(\alpha,\beta)\phi_j(\beta)\,d\beta = \lambda_j\phi_j(\alpha)}\,.$$

The $\{\phi_i\}$ which allow the representation of $X(t)$ to be a sequence of uncorrelated random variables must satisfy this integral equation. This type of equation occurs often in applied mathematics; it is termed the *eigenequation*. The sequences $\{\phi_i\}$ and $\{\lambda_i\}$ are the eigenfunctions and eigenvalues of $K_X(\alpha,\beta)$, the covariance function of $X(t)$. It is easily verified that:

$$K_X(t,u) = \sum_{i=1}^{\infty}\lambda_i\phi_i(t)\phi_i(u)$$

This result is termed *Mercer's Theorem*.

The approach to solving for the eigenfunctions and eigenvalues of $K_X(t,u)$ is to convert the integral equation into an ordinary differential equation which can be solved. This approach is best illustrated by an example.

Example

$K_X(t,u)=\sigma^2\min(t,u)$. The eigenequation can be written in this case as

$$\sigma^2\left[\int_0^t u\phi(u)\,du + t\int_t^T \phi(u)\,du\right] = \lambda\phi(t)\,.$$

Evaluating the first derivative of this expression,

$$\sigma^2 t\phi(t) + \sigma^2 \int_t^T \phi(u)\,du - \sigma^2 t\phi(t) = \lambda \frac{d\phi(t)}{dt}$$
$$\text{or} \quad \sigma^2 \int_t^T \phi(u)du = \lambda \frac{d\phi}{dt}\,.$$

Evaluating the derivative of the last expression yields the simple equation

$$-\sigma^2\phi(t) = \lambda \frac{d^2\phi}{dt^2}\,.$$

This equation has a general solution of the form $\phi(t) = A\sin\frac{\sigma}{\sqrt{\lambda}}t + B\cos\frac{\sigma}{\sqrt{\lambda}}t$. It is easily seen that B must be zero. The amplitude A is found by requiring $\|\phi\| = 1$. To find λ, one must return to the original integral equation. Substituting, we have

$$\sigma^2 A\int_0^t u\sin\frac{\sigma}{\sqrt{\lambda}}u\,du + \sigma^2 tA\int_t^T \sin\frac{\sigma}{\sqrt{\lambda}}u\,du = \lambda A\sin\frac{\sigma}{\sqrt{\lambda}}t$$

After some manipulation, we find that

$$A\lambda\sin\frac{\sigma}{\sqrt{\lambda}}t - A\sigma t\sqrt{\lambda}\cos\frac{\sigma}{\sqrt{\lambda}}T = \lambda A\sin\frac{\sigma}{\sqrt{\lambda}}t \quad \forall t\in[0,T)\,.$$
$$\text{or} \quad A\sigma t\sqrt{\lambda}\cos\frac{\sigma}{\sqrt{\lambda}}T = 0 \quad \forall t\in[0,T)\,.$$

Therefore, the cosine term must be zero, which means that we must require $\frac{\sigma}{\sqrt{\lambda}}T = (n-1/2)\pi$, $n = 1,2,\ldots$, or

$$\lambda_n = \frac{\sigma^2T^2}{(n-1/2)^2\pi^2}$$
$$\Longrightarrow\ \phi_n(t) = \left(\frac{2}{T}\right)^{1/2}\sin\frac{(n-1/2)\pi t}{T}\,.$$

The Karhunen-Loève expansion has several important properties.

- The eigenfunctions of a positive-definite covariance function constitute a complete set. One can easily show that these eigenfunctions are also mutually orthogonal with respect to both the usual inner product and with respect to the inner product derived from the covariance function (see problem 2.47).
- If $X(t)$ Gaussian, X_i are Gaussian random variables. As the random variables $\{X_i\}$ are uncorrelated and Gaussian, the $\{X_i\}$ comprise a sequence of statistically independent Gaussian random variables.
- Assume $K_X(t,u) = \frac{N_0}{2}\delta(t-u)$: the stochastic process $X(t)$ is white. Then

$$\int \frac{N_0}{2}\delta(t-u)\phi(u)du = \lambda\phi(t)$$

for all $\phi(t)$. Consequently, if $\lambda_i = N_0/2$, this constraint equation is satisfied *no matter what choice is made for the orthonormal set* $\{\phi_i(t)\}$. Therefore, the representation of white, Gaussian processes consists of a sequence of statistically independent, identically-distributed (mean zero and variance $N_0/2$) Gaussian random variables. This example constitutes the simplest case of the Karhunen-Loève expansion.

Problems

2.1 Space Exploration and MTV

Joe is an astronaut for project Pluto. The mission success or failure depends only on the behavior of three major systems. Joe feels that the following assumptions are valid and apply to the performance of the entire mission:

- The mission is a failure only if two or more major systems fail.
- System I, the Gronk system, fails with probability 0.1.
- System II, the Frab system, fails with probability 0.5 if at least one other system fails. If no other system fails, the probability the Frab system fails is 0.1.
- System III, the beer cooler (obviously, the most important), fails with probability 0.5 if the Gronk system fails. Otherwise the beer cooler cannot fail.

(a) What is the probability that the mission succeeds but that the beer cooler fails?

(b) What is the probability that all three systems fail?

(c) Given that more than one system failed, determine the probability that:

(i) The Gronk did not fail.

(ii) The beer cooler failed.

(iii) Both the Gronk and the Frab failed.

(d) About the time Joe was due back on Earth, you overhear a radio broadcast about Joe while watching MTV. You are not positive what the radio announcer said, but you decide that it is twice as likely that you heard "mission a success" as opposed to "mission a failure". What is the probability that the Gronk failed?

2.2 Probability Density Functions?

Which of the following are probability density functions? Indicate your reasoning. For those that are valid, what is the mean and variance of the random variable?

(i) $p_X(x) = \dfrac{e^{-|x|}}{2}$ (ii) $p_X(x) = \dfrac{\sin 2\pi x}{\pi x}$

(iii) $p_X(x) = \begin{cases} 1-|x| & |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$ (iv) $p_X(x) = \begin{cases} 1 & |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$

(v) $p_X(x) = \dfrac{1}{4}\delta(x+1) + \dfrac{1}{2}\delta(x) + \dfrac{1}{4}\delta(x-1)$ (vi) $p_X(x) = \begin{cases} e^{-(x-1)} & x \ge 1 \\ 0 & \text{otherwise} \end{cases}$

2.3 Gaussian Random Vectors

The probability density function for jointly Gaussian random vectors is has the form

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{|2\pi\mathbf{K}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\right\}$$

Here, $|2\pi\mathbf{K}|$ denotes the determinant of $2\pi\mathbf{K}$. This joint distribution is important because it is the limiting density of a sum of statistically independent, identically distributed random vectors.* Several properties of jointly Gaussian random variables need to be derived. The derivations are most easily done using characteristic functions.

(a) The characteristic function of a random vector is defined to be

$$\Phi_{\mathbf{X}}(j\boldsymbol{\nu}) = \int p_{\mathbf{X}}(\mathbf{x}) e^{j\boldsymbol{\nu}^t\mathbf{x}} d\mathbf{x}$$

Show that the characteristic function of the jointly Gaussian density equals

$$\Phi_{\mathbf{X}}(j\boldsymbol{\nu}) = \exp\left\{j\boldsymbol{\nu}^t\mathbf{m} - \frac{1}{2}\boldsymbol{\nu}^t\mathbf{K}\boldsymbol{\nu}\right\}$$

*Suitable conditions are required just as in the Central Limit Theorem for random variables (see §2.1.10).

(b) Show that $\mathsf{E}[\mathbf{X}] = \mathbf{m}$ and that $\mathsf{E}[\mathbf{X}\mathbf{X}^t] = \mathbf{K} + \mathbf{m}\mathbf{m}^t$. In other words, the quantity $\mathbf{m}$ in the formula for the jointly Gaussian density is indeed the mean vector and quantity $\mathbf{K}$ is the covariance matrix.

(c) Show that all the marginal densities that can be derived from the jointly Gaussian density are indeed Gaussian.

(d) Joint densities exist that have Gaussian marginals but don't have the form of the jointly Gaussian density. Show that the following joint density is indeed a density (but not a jointly Gaussian density) and that its marginals are Gaussian.

$$p_{X_1,X_2}(x_1,x_2) = \begin{cases} \frac{1}{\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2}+\frac{x_2^2}{\sigma_2^2}\right)\right\} & x_1 \cdot x_2 \geq 0 \\ 0 & x_1 \cdot x_2 < 0 \end{cases}$$

2.4 Generating Random Variables

A crucial skill in developing simulations of communication systems is *random variable generation*. Most computers (and environments like MATLAB) have software that generates statistically independent, uniformly distributed, random sequences. In MATLAB, the function is `rand`. We want to change the probability distribution to one required by the problem at hand. One technique is known as the *distribution method*.

(a) If $P_X(x)$ is the desired distribution, show that $U = P_X(X)$ (applying the distribution function to a random variable having that distribution) is uniformly distributed over $[0,1)$.

This result means that the random variable resulting from applying the inverse of the distribution function to a uniformly distributed random variable $U \sim [0,1] - P_X^{-1}(U)$ — has the probability distribution $P_X(x)$. Consequently, to generate a random variable having any distribution we want, we only need the inverse function of the distribution function.

(b) Why is the Gaussian not in the class of "nice" probability distribution functions?

(c) How would you generate random variables having the Laplacian, the Cauchy and the hyperbolic secant $p_X(x) = (1/2)\mathrm{sech}(\pi x/2)$ densities?

(d) Write MATLAB functions that generate these random variables. Again use `hist` to plot the probability function. What do you notice about these random variables?

2.5 Cauchy Random Variables

The random variables X_1 and X_2 have the joint pdf

$$p_{X_1,X_2}(x_1,x_2) = \frac{1}{\pi^2}\frac{b_1 b_2}{(b_1^2+x_1^2)(b_2^2+x_2^2)}\ , b_1, b_2 > 0\ .$$

(a) Show that X_1 and X_2 are statistically independent random variables with Cauchy density functions.

(b) Show that $\Phi_{X_1}(j\nu) = e^{-b_1|\nu|}$.

(c) Define $Y = X_1 + X_2$. Determine $p_Y(y)$.

(d) Let $\{Z_i\}$ be a set of N statistically independent Cauchy random variables with $b_i = b$, $i = 1,\ldots,N$. Define

$$Z = \frac{1}{N}\sum_{i=1}^{N} Z_i.$$

Determine $p_Z(z)$. Is Z—the sample mean—a good estimate of the expected value $\mathsf{E}[Z_i]$?

2.6 The Correlation Coefficient

The random variables X, Y have the joint probability density $p_{X,Y}(x,y)$. The *correlation coefficient* $\rho_{X,Y}$ is defined to be

$$\rho_{X,Y} \equiv \frac{\mathsf{E}\left[(X-m_X)(Y-m_y)\right]}{\sigma_X\sigma_Y}\ .$$

(a) Using the Cauchy-Schwarz inequality, show that correlation coefficients always have a magnitude less than to equal to one.

(b) We would like find an affine estimate of one random variable's value from the other. So, if we wanted to estimate X from Y, our estimate $\widehat{X}$ has the form $\widehat{X} = aY + b$, where a, b are constants to be found. Our criterion is the mean-squared estimation error: $\varepsilon^2 = \mathsf{E}\left[(\widehat{X} - X)^2\right]$. First of all, let $a = 0$: we want to estimate X without using Y at all. Find the optimal value of b.

(c) Find the optimal values for both constants. Express your result using the correlation coefficient.

(d) What is the expected value of your estimate?

(e) What is the smallest possible mean-squared error? What influence does the correlation coefficient have on the estimate's accuracy?

2.7 Order Statistics

Let $X_1, \ldots, X_N$ be independent, identically distributed random variables. The density of each random variable is $p_X(x)$. The *order statistics* $X_0(1), ..., X_0(N)$ of this set of random variables is the set that results when the original one is ordered (sorted).

$$X_0(1) \leq X_0(2) \leq ... \leq X_0(N)$$

(a) What is the joint density of the original set of random variables?

(b) What is the density of $X_0(N)$, the largest of the set?

(c) Show that the joint density of the ordered random variables is

$$p_{X_0(1),...,X_0(N)}(x_1, ..., x_N) = N! p_X(x_1) \cdots p_X(x_N)$$

(d) Consider a Poisson process having constant intensity λ_0. N events are observed to occur in the interval $[0, T)$. Show that the joint density of the times of occurrence $W_1, \ldots, W_N$ is the same as the order statistics of a set of random variables. Find the common density of these random variables.

2.8 Estimating Characteristic Functions

Suppose you have a sequence of statistically independent, identically distributed random variables $X_1, \ldots, X_N$. From these, we want to estimate the characteristic function of the underlying random variable. One way to estimate it is to compute

$$\widehat{\Phi}_X(j\nu) = \frac{1}{N} \sum_{n=1}^{N} e^{j\nu X_n}$$

(a) What is the expected value of the estimate?

(b) Does this estimate converge to the actual characteristic function? There are many criteria for "convergence." For this problem, let's use mean-squared error: does the mean-squared error go to zero for large N? If the estimate converges in the mean-squared sense, demonstrate how; if not, why not?

2.9 Kullback-Leibler Distance

Calculate the Kullback-Leibler distance between the following pairs of densities. Use these results to find the Fisher information for the mean parameter m.

(a) Jointly Gaussian random vectors having the same covariance matrix but dissimilar mean vectors.

(b) Two Poisson random variables having average rates λ_0 and λ_1. In this example, the observation time T plays the role of the number of observations.

(c) Two sequences of statistically independent Laplacian random variables having the same variance but different means.

(d) Deduce from these expressions what the Cramér-Rao bound is for the relevant example in each case.

(e) Plot the Kullback-Leibler distances for the Laplacian case and for the Gaussian case of statistically independent random variables. Set the variance equal to σ^2 in each case and plot the distances as a function of m/σ.

2.10 Nonlinearities and Processes

The Gaussian wide-sense stationary random process X_t, having mean zero and correlation function $R_X(\tau)$, is squared during signal processing: $Y_t = X_t^2$.

(a) Is Y_t a wide-sense stationary process? Why or why not?

(b) What is the probability density of Y_t?

(c) If it exists, find the power density spectrum of Y_t.

2.11 Simple Processes

Determine the mean and correlation function of each of the processes X_t defined below.

(a) X_t is defined by the following equally likely sample functions.

$$X_t(\omega_1) = 1 \qquad X_t(\omega_3) = \sin \pi t$$
$$X_t(\omega_2) = -2 \qquad X_t(\omega_4) = \cos \pi t$$

(b) X_t is defined by $X_t = \cos(At + \theta)$, where A and θ are statistically independent random variables. θ is uniformly distributed over $[0, 2\pi)$ and A has the density function

$$p_A(A) = \frac{1}{\pi(1+A^2)}$$

2.12 An Elementary Process

The joint density of the amplitudes of a stochastic process X_t at the specific times $t = t_1$ and $t = t_2$ $(t_2 > t_1)$ is found to be

$$p_{X_{t_1},X_{t_2}}(x_1, x_2) = \begin{cases} \text{constant} & x_1 > x_2\ , 0 < x_1, x_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

This joint density is found to be a valid joint density for X_t and X_u when $|t-u| = |t_2 - t_1|$.

(a) Find the correlation function $R_X(t,u)$ at the times $t = t_1$ and $u = t_2$.

(b) Find the expected value of X_t for all t.

(c) Is this process wide-sense stationary?

2.13 Correlation Functions and Power Spectra

(a) Which of the following are valid correlation functions? Indicate your reasoning.

(i) $R_X(\tau) = e^{-|\tau|} - e^{-2|\tau|}$ (ii) $R_X(\tau) = \dfrac{5 \sin 1000\tau}{\tau}$

(iii) $R_X(\tau) = \begin{cases} 1 - \frac{|\tau|}{T} & |\tau| \le T \\ 0 & \text{otherwise} \end{cases}$ (iv) $R_X(\tau) = \begin{cases} 1 & |\tau| \le T \\ 0 & \text{otherwise} \end{cases}$

(v) $R_X(\tau) = \delta(\tau) + 25$ (vi) $R_X(\tau) = \delta(\tau+1) + \delta(\tau) + \delta(\tau-1)$

(b) Which of the following are valid power density spectra? Indicate your reasoning.

(i) $\mathscr{S}_X(f) = \dfrac{\sin \pi f}{\pi f}$ (ii) $\mathscr{S}_X(f) = \left(\dfrac{\sin \pi f}{\pi f}\right)^2$

(iii) $\mathscr{S}_X(f) = \exp\left\{-\dfrac{(f-f_0)^2}{4}\right\}$ (iv) $\mathscr{S}_X(f) = e^{-|f|} - e^{-2|f|}$

(v) $\mathscr{S}_X(f) = 1 + 0.25e^{-j2\pi f}$ (vi) $\mathscr{S}_X(f) = \begin{cases} 1 & |f| \le 1/T \\ 0 & \text{otherwise} \end{cases}$

2.14 Sampling Stochastic Processes

Let X_t be a wide-sense stationary process bandlimited to W Hz. The sampling interval T_s satisfies $T_s \le \frac{1}{2W}$.

(a) What is the covariance of successive samples?

(b) Now let X_t be Gaussian. What conditions on T_s will insure that successive samples will be statistically independent?

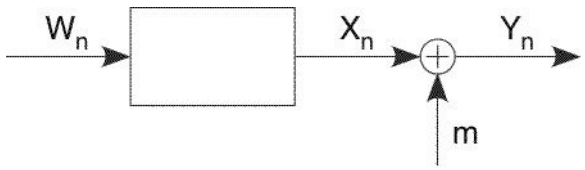
(c) Now assume the process is *not* strictly bandmilited to W Hz. This process serves as the input to an ideal lowpass filter having cutoff frequency W to produce the process Y_t. This output is sampled every $\frac{1}{2W}$ seconds, then converted back to an analog waveform to yield an approximate representation Z_t of the original signal X_t. Show that the mean-squared value of the sampling error, defined to be $\varepsilon^2 = \mathsf{E}[(X_t - Z_t)^2]$, is given by $\varepsilon^2 = 2\int_W^\infty \mathscr{S}_x(f)\,df$.

2.15 Estimating the Mean

Suppose you have stochastic process Y_n produced by the depicted system. The input W_n is discrete-time white noise (not necessarily Gaussian) having zero mean and correlation function $R_W(l) = \sigma_W^2\delta(l)$. The system relating X_n to the white-noise input is governed by the difference equation

$$X_n = aX_{n-1} + W_n, \quad |a| < 1\ .$$

The quantity m is an unknown constant.



(a) Is the process Y_n stationary? Why or why not?

(b) What is the correlation function of Y_n?

(c) We want to estimate the constant m by averaging Y_n over a finite time interval: $\widehat{m} = \frac{1}{N}\sum_{n=0}^{N-1} Y_n$. What is the expected value and variance of this estimate?

2.16 Estimating a Process's Amplitude Distribution

Let X_n be a strict-sense stationary discrete-time random process. To estimate the process's amplitude distribution, we define a set of amplitude intervals A_i and compute the number of times the values of the process falls within the interval. Using the indicator function notation,

$$I_A(X_n) = \begin{cases} 1 & X_n \in A \\ 0 & X_n \notin A \end{cases}$$

the estimate of $\Pr[X_n \in A]$ is

$$\widehat{\Pr}[X_n \in A] = \frac{1}{N}\sum_{n=0}^{N-1} I_A(X_n)\ .$$

(a) What is the probability distribution of this estimate when the process is white noise?

(b) Under what conditions does this estimate converge in the mean-square sense?

(c) What is the correlation coefficient between the estimates $\widehat{\Pr}[X_n \in A_i]$ and $\widehat{\Pr}[X_n \in A_j]$, $A_i \cap A_j = \emptyset$, made from the same set of N observations?

2.17 A Non-Gaussian Process

Let $\{X_l\}$ denote a sequence of independent, identically distributed random variables. This sequence serves as the input to a discrete-time system having an input-output relationship given by the difference equation

$$Y_l = aY_{l-1} + X_l$$

(a) If $X_l \sim \mathscr{N}(0,\sigma^2)$, find the probability density function of each element of the output sequence $\{Y_l\}$.

(b) Show that $|\Phi_{X_l}(j\nu)| \leq \Phi_{X_l}(j0)$ for all choices of ν no matter what the amplitude distribution of X_l may be.

(c) If X_l is non-Gaussian, the computation of the probability density of Y_l can be difficult. On the other hand, if the density of Y_l is known, the density of X_l can be found. How is the characteristic function of X_l related to the characteristic function of Y_l?

(d) Show that if Y_l is uniformly distributed over $[-1,1)$, the only allowed values of the parameter a are those equalling $1/m$, $m = \pm 2, \pm 3, \pm 4, \ldots$.

2.18 Random-Phase Sinusoids

A random-phase sinusoid is a stochastic process X_t defined to be $X_t = \cos(2\pi f_0 t + \theta)$, where θ is a random variable having some probability density function $p_\theta(\theta)$ defined over $0 \leq \theta < 2\pi$. θ is not necessarily uniformly distributed.

(a) Show that necessary and sufficient conditions for the stochastic process X_t to be wide-sense stationary is that the characteristic function $\Phi_\theta(j\nu)$ satisfy

$$\Phi_\theta(j1) = 0 = \Phi_\theta(j2).$$

NOTE: f_0 is a constant and is not a random variable.

(b) Let $X_t = A\cos(2\pi f_0 t) + B\sin(2\pi f_0 t)$, where A and B are random variables and f_0 constant. Find necessary and sufficient conditions for X_t to be wide-sense stationary.

(c) Consider the process defined by:

$$X_t = A\cos(2\pi f_0 t)$$

where A is a random variable and f_0 is a constant. Under what conditions is X_t wide-sense stationary?

2.19 Properties of Correlation Functions

(a) Show that correlation and covariance functions have the following properties:

1. $R_X(t,u) = R_X(u,t)$
2. $R_X(\tau) = R_X(-\tau)$
3. $K_X^2(t,u) \leq K_X(t,t) \cdot K_X(u,u)$
4. $|K_X(t,u)| \leq \frac{1}{2}[K_X(t,t) + K_X(u,u)]$
5. $|R_X(\tau)| \leq R_X(0)$

(b) Let X_t be a wide-sense stationary random process. If $s(t)$ is a deterministic function and we define $Y_t = X_t + s(t)$, what is the expected value and correlation function of Y_t?

(c) Which of the following are valid correlation functions? Justify your answer.

1. $R_X(t,u) = e^{-(t-u)^2}$
2. $R_X(t,u) = \sigma^2 \max(t,u)$
3. $R_X(t,u) = e^{-(t-u)}$
4. $R_X(t,u) = \cos t \cdot \cos u$

2.20 Models for Process Generation

It is desired to generate a wide-sense stationary process with correlation function

$$R_X(\tau) = e^{-|\tau|}.$$

Two methods are proposed.

1. Let $X_t = A\cos(2\pi F t + \theta)$ where A, F, and θ are statistically independent random variables.
2. Define X_t by:

$$X_t = \int_0^\infty h(\alpha) N_{t-\alpha}\, d\alpha$$

where N_t is white and $h(t)$ is the impulse response of the appropriate filter.

(a) Find at least two impulse responses $h(t)$ that will work in method 2.

(b) Specify the densities for A, F, θ in method 1 that yield the desired results.

(c) Sketch sample functions generated by each method. Interpret your result. What are the technical differences between these processes?

2.21 One Model Fits all?

Let the stochastic process X_t be defined by

$$X_t = A\cos(2\pi Ft + \theta)$$

where A, F, and θ are statistically independent random variables. The random variable θ is uniformly distributed over the interval $[-\pi, \pi)$. The densities of the other random variables are to be determined.

(a) Show that X_t is a wide-sense stationary process.

(b) Is X_t strict-sense stationary? Why or why not?

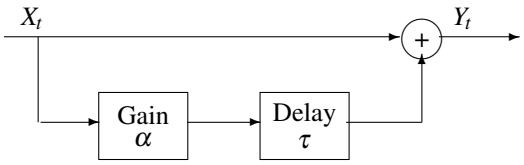
(c) The inventor of this process claims that X_t can have *any* correlation function one desires by manipulating the densities of A and F. Demonstrate the validity of this result and the requirements these densities must satisfy.

(d) The inventor also claims that X_t can have *any* first-order density one desires so long as the desired density is bounded. Show that this claim is also valid. Furthermore, show that the requirements placed on the densities of A and F are consistent with those found in the previous part.

(e) Could this process be fruitfully used in simulations to emulate a process having a specific correlation function and first-order density? In other words, would statistics computed from the simulation results be meaningful? Why or why not?

2.22 Multipath Models

Let X_t be a Gaussian random process with mean $m_X(t)$ and covariance function $K_X(t,u)$. The process is passed through the depicted system.



(a) Is Y_t a Gaussian process? If so, compute the pdf of Y_t.

(b) What are the mean and covariance functions of Y_t?

(c) If X_t is stationary, is Y_t stationary?

(d) Compute the cross-correlation function between X_t and Y_t.

2.23 Nonlinear Processing of a Random Process

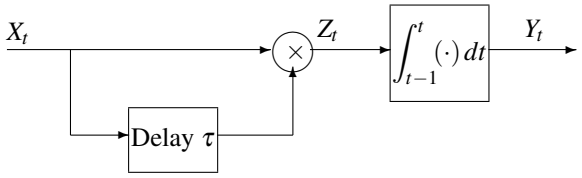
A stochastic process X_t is defined to be

$$X_t = \cos(2\pi Ft + \theta)$$

where F and θ are statistically independent random variables. The quantity θ is uniformly distributed over $[-\pi, \pi)$ and F can assume one of the values 1, 2, or 3 with equal probability.

(a) Compute the mean and correlation function of X_t.

This process serves as the input to the following system.



Note: The signals are *multiplied*, not summed, at the node located just before the integrator. Y_t and Z_t are related by

$$Y_t = \int_{t-1}^{t} Z_\alpha \, d\alpha$$

(b) Is the process Z_t wide-sense stationary?

(c) What is the expected value and correlation function of Y_t?

2.24 Processes and Differential Equations

The white process X_t serves as the input to a system having output Y_t. The input-output relationship of this system is determined by the differential equation

$$\dot{Y}_t + 2Y_t = X_t$$

(a) Find the mean and correlation function of Y_t.

(b) Compute the cross-correlation function between X_t and Y_t.

(c) Show that the correlation function of Y_t obeys a homogeneous version of the differential equation governing the system for positive values of τ.

$$\dot{R}_Y(\tau) + 2R_Y(\tau) = 0 \, , \tau > 0$$

Do **not** use your answer to part (a) to work this part. Rather, show the validity of this result in a more general fashion.

2.25 Joint Statistics of a Process and its Derivative

Let X_t be a wide-sense stationary stochastic process. Let $\dot{X}_t$ denote the derivative of X_t.

(a) Compute the expected value and correlation function of $\dot{X}_t$ in terms of the expected value and correlation function of X_t.

(b) Under what conditions are $\dot{X}_t$ and X_t orthogonal? In other words, when does $\langle \dot{X}, X \rangle = 0$ where $\langle \dot{X}, X \rangle = \mathsf{E}[\dot{X}_t X_t]$?

(c) Compute the mean and correlation function of $Y_t = X_t - \dot{X}_t$.

(d) The bandwidth of the process X_t can be defined by

$$B_X^2 = \frac{\displaystyle\int_{-\infty}^{\infty} f^2 \mathscr{S}_X(f)\, df}{\displaystyle\int_{-\infty}^{\infty} \mathscr{S}_X(f)\, df}$$

Express this definition in terms of the mean and correlation functions of X_t and $\dot{X}_t$.

(e) The statistic U is used to count the average number of excursions of the stochastic process X_t across the level $X_t = A$ in the interval $[0, T]$. One form of this statistic is

$$U = \frac{1}{T} \int_0^T \left| \frac{d}{dt} \mathrm{u}(X_t - A) \right| dt$$

where $\mathrm{u}(\cdot)$ denotes the unit step function. Find the expected value of U, using in your final expression the formula for B_X. Assume that the conditions found in part (b) are met and X_t is a Gaussian process.

2.26 Independent Increment Processes

A stochastic process X_t is said to have *stationary, independent increments* if, for $t_1 < t_2 < t_3 < t_4$:

- The random variable $X_{t_2} - X_{t_1}$ is statistically independent of the random variable $X_{t_4} - X_{t_3}$.
- The pdf of $X_{t_2} - X_{t_1}$ is equal to the pdf of $X_{t_2+T} - X_{t_1+T}$ for all t_1, t_2, T.

The process is identically equal to zero at $t = 0$: $\Pr[X_0 = 0] = 1$.

(a) What is the expected value and variance of $X_{t_1+t_2}$?
Hint: Write $X_{t_1+t_2} = [X_{t_1+t_2} - X_{t_1}] + [X_{t_1} - X_0]$.

(b) Using the result of part (a), find expressions for $\mathsf{E}[X_t]$ and $\mathsf{var}[X_t]$.

(c) Define $\Phi_{X_t}(j\nu)$ to be the characteristic function of the first-order density of the process X_t. Show that this characteristic function must be of the form:

$$\Phi_{X_t}(j\nu) = e^{t \cdot f(\nu)}$$

where $f(\nu)$ is a conjugate-symmetric function of ν.

(d) Compute $K_X(t,u)$.

(e) The process X_t is passed through a linear, time-invariant filter having the transfer function $H(f) = j2\pi f$. Letting Y_t denote the output, determine $K_Y(t,u)$.

2.27 Poisson Processes

Let N_t be a Poisson process with intensity $\lambda(t)$.

(a) What is the expected value and variance of the number of events occurring in the time interval $[t,u)$?

(b) Under what conditions is N_t a stationary, independent increment process?

(c) Compute $R_N(t,u)$.

(d) Assume that $\lambda(t) = \lambda_0$, a constant. What is the conditional density $p_{W_n|W_{n-1}}(w_n|w_{n-1})$? From this relationship, find the density of τ_n, the time interval between W_n and W_{n-1}.

2.28 Optical Communications

In optical communication systems, a photomultiplier tube is used to convert the arrival of photons into electric pulses so that each arrival can be counted by other electronics. Being overly clever, a clever Rice engineer bought a photomultiplier tube from AGGIE PMT, Inc. The AGGIE PMT device is unreliable. When it is working, each photon is properly converted to an electric pulse. When not working, it has "dead-time" effects: the conversion of a photon arrival blocks out the conversion of the next photon. After a photon arrival has been missed, the device converts the next arrival properly. To detect whether the Aggie device is working properly or not, the clever Rice engineer decides to use results from a statistical signal processing course he is taking to help him. A calibrated light source is used to give an average arrival rate of λ photons/sec on the surface of the photomultiplier tube. Photon arrivals are described by a Poisson process.

(a) Find the density of the time between electric pulses if the AGGIE device has these dead-time effects.

(b) Can the times of occurrence of electric pulses be well-described by a Poisson process when the dead-time effects are present? If so, find the parameters of the process; if not, state why.

(c) Assuming the device is as likely to not be working as it is to be working, find a procedure to determine its mode of operation based on the observation of the time between two successive electric pulses.

2.29 Warping the Time Axis

In this problem, assume that the process X_t has stationary, independent increments.

(a) Define Y_t to be:

$$Y_t = g(t) \cdot X_{h(t)/g(t)}$$

where $g(t)$ and $h(t)$ are deterministic functions and $h(t)/g(t)$ is a strictly increasing function. Find the mean and covariance functions of Y_t.

(b) A stochastic process X_t is said to be a *Markov* process if, for $t_1 < t_2 < \ldots < t_{n-1} < t_n$, the conditional density of X_{t_n} satisfies:

$$p_{X_{t_n}|X_{t_1},\ldots,X_{t_{n-1}}}(X_n|X_1,\ldots,X_{n-1}) = p_{X_{t_n}|X_{t_{n-1}}}(X_n|X_{n-1}).$$

Show that all independent-increment processes are Markov processes.

2.30 Mean-Square Continuity

Let X_t be a stochastic process with correlation function $R_X(t,u)$. X_t is said to be *mean-square continuous* if

$$\lim_{t\to u} \mathsf{E}[(X_t - X_u)^2] = 0 \ , \ \text{forall } t,u \ .$$

(a) Show that X_t is mean-square continuous if and only if the correlation function $R_X(t,u)$ is continuous at $t = u$.

(b) Show that if $R_X(t,u)$ is continuous at $t = u$, it is continuous for all t and u.

(c) Show that a zero-mean, independent-increment process with stationary increments is mean-square continuous.

(d) Show that a stationary Poisson process is mean-square continuous. Note that this process has no continuous sample functions, but is continuous in the mean-square sense.

2.31 The Bispectrum

The idea of a correlation function can be extended to higher order moments. For example, the third order "correlation" function $R_X^{(3)}(t_1,t_2,t_3)$ of a random process X_t is defined to be

$$R_X^{(3)}(t_1,t_2,t_3) = \mathsf{E}[X_{t_1}X_{t_2}X_{t_3}]$$

(a) Show that if X_t is strict-sense stationary, then the third-order correlation function depends only on the time differences $t_2 - t_1$ and $t_3 - t_1$.

(b) Find the third-order correlation function of $X_t = A\cos(2\pi f_0 t + \Theta)$, where $\Theta \sim U[-\pi,\pi)$ and A, f_0 are constants.

(c) Let $Z_t = X_t + Y_t$, where X_t is Gaussian, Y_t is non-Gaussian, and X_t, Y_t are statistically independent, zero-mean processes. Find the third-order correlation function of Z_t.

2.32 Martingales

A sequence of random variables $X_0, X_1, \ldots, X_l, \ldots$ is said to be a *martingale* if the conditional expectation of the "present" value given the entire past satisfies

$$\mathsf{E}[X_l|X_{l-1},X_{l-2},\ldots,X_0] = X_{l-1} \ .$$

(a) Show that the mean of a martingale is a constant. If X_l is a zero-mean martingale having a possibly time-varying variance $\sigma^2(l)$, show that its correlation function is given by

$$R_X(k,l) = \sigma^2\big(\min(k,l)\big)$$

(b) Let $X_l = \sum_{k=0}^{l} W_k$, where $W_k = \pm 1$ with equal probability and the sequence $\{W_k\}$ is mean-square independent: $\mathsf{E}[W_k|W_{k-1},\ldots,W_0] = \mathsf{E}[W_k]$. Show that X_l is a martingale.

(c) An *independent increments* sequence has the property that first differences are independent random variables. Thus, for X_l to have independent increments, $X_l - X_{l-1}$ is statistically independent of $X_{l-1} - X_{l-2}$. Show that if X_l has independent increments, then $X_l - \mathsf{E}[X_l]$ is a martingale. Are all martingales independent increment processes?

(d) The *likelihood ratio* is a very important quantity in statistical hypothesis testing and in detection theory. In such problems, we are trying to decide if the data $X_0,\ldots,X_l = \mathbf{X}$ were produced by probability law $p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)$ or $p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)$, where $\mathscr{M}_0$ and $\mathscr{M}_1$ describe alternative models for the data generation process. The likelihood ratio is defined to be

$$\Lambda_L(\mathbf{X}) \equiv \frac{p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)}{p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)} \ .$$

Show that under model $\mathscr{M}_0$, $\Lambda_L(\mathbf{X})$ is a martingale.

2.33 Shot Noise

Shot noise is noise measured in vacuum tube circuits which is due to the spontaneous emission of electrons from each tube's cathode. The electron emission is assumed to be described as a stationary Poisson process of intensity λ. The impact of each electron on a tube's anode causes a current to flow in the attached circuit equal to the impulse response of the circuit. Thus, shot noise is often modeled as a sequence of impulses (whose times of occurrence are a Poisson process) passing through a linear, time-invariant system.

(a) What is the correlation function of X_t?
Hint: Relate X_t to the counting process N_t.

(b) Show that for any wide-sense stationary process X_t for which $\lim_{\tau\to\infty} R_X(\tau) = 0$, the mean of X_t is zero. Use this result to show that if $\lim_{\tau\to\infty} R_X(\tau)$ exists, the value of the limit equals the square of the mean of the process.

(c) Find the power density spectrum of the shot noise process Y_t.

(d) Evaluate the mean and variance of Y_t.

2.34 Random Telegraph Wave

One form of the *random telegraph wave* X_t is derived from a stationary Poisson process $N_{0,t}$ having constant event rate λ.

$$X_t = \begin{cases} +1 & N_{0,t} \text{ even} \\ -1 & N_{0,t} \text{ odd} \end{cases}$$

For $t < 0$, the process is undefined. $N_{0,t}$ denotes the number of events that have occurred in the interval $[0,t)$ and has a probability distribution given by

$$\Pr[N_{0,t} = n] = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad t \geq 0 \,.$$

Note that $\Pr[N_{0,0} = 0] \equiv 1$.

(a) What is the probability distribution of X_t?

(b) Find the mean and correlation function of X_t.

(c) Is this process wide-sense stationary? stationary in the stricter sense? Provide your reasoning.

2.35 Suppressing Noise

Noise reduction filters are used to reduce, as much as possible, the noise component of a noise-corrupted signal. Let the signal of interest be described as a wide-sense stationary process X_t. The observed signal is given by $Y_t = X_t + N_t$, where N_t is a process modeling the noise that is statistically independent of the signal.

(a) Assuming that the noise is white, find a relationship that the transfer function of the filter must satisfy to maximize the signal-to-noise ratio (i.e., the ratio of the signal power to the noise power) in the filtered output.

(b) Compute the resulting signal-to-noise ratio when the signal correlation function is

$$R_X(\tau) = \sigma^2 e^{-a|\tau|}.$$

2.36 Analog Communication

The message M_t is to be transmitted using amplitude modulation. The transmitted signal has the form

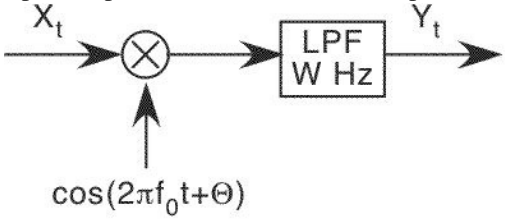
$$X_t = (1 + M_t)\cos(2\pi f_0 t + \Theta),\ f_0 = \text{constant}, \Theta \sim U(-\pi, \pi]$$

The message is a wide-sense stationary Gaussian process having zero mean. It is bandlimited to W Hz and statistically independent of the random phase Θ. A few technical details are not quite worked out...

(a) For technical reasons, the transmitter needs to *clip* the message so that it never exceeds 1 in magnitude. What is the probability that the magnitude of M_t at a randomly chosen time exceeds 1? Express your answer in terms of the mean and correlation function of M_t.

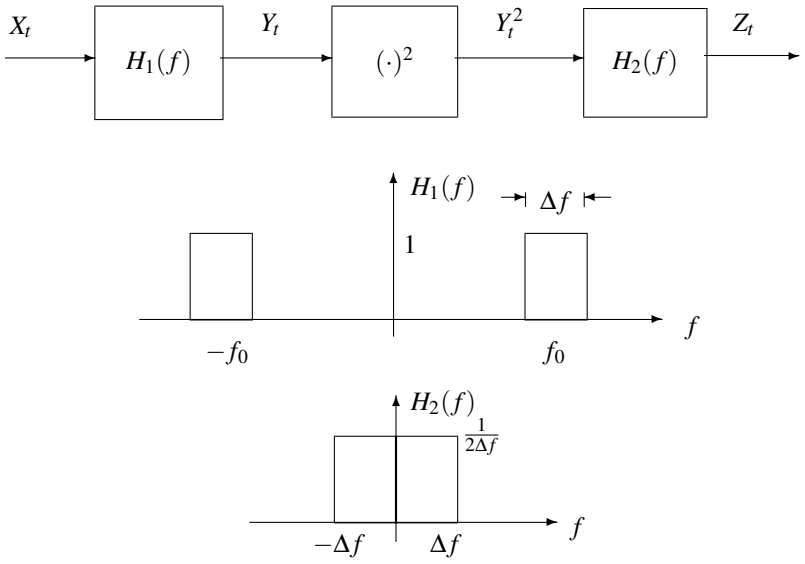
(b) What is the power spectrum of X_t? Assume that the probability of clipping the message is negligible.

(c) A standard receiver multiplies the incoming signal by a sinusoid having a frequency and phase equal to that of the carrier, then passes the product through an ideal lowpass filter having a cutoff frequency of W Hz. What is the power spectrum of the receiver's output? Assume $f_0 \gg W$.



2.37 Measuring Power Spectra

In practice, one often wants to measure the power density of a stochastic process. For the purposes of this problem, assume the process X_t is wide-sense stationary, zero mean, and Gaussian. The following measurement system is proposed.



where $H_1(f)$ is the transfer function of an ideal bandpass filter and $H_2(f)$ is an ideal lowpass. Assume that Δf is small compared to range of frequencies over which $\mathscr{S}_X(f)$ varies.

(a) Find the mean and correlation function of Y_t^2 in terms of the second-order statistics of X_t.

(b) Compute the power density spectrum of the process Z_t.

(c) Compute the expected value of Z_t.

(d) By considering the variance of Z_t, comment on the accuracy of this measurement of the power density of the process X_t.

2.38 Chemistry Experiments

A student in freshman chemistry lab is frustrated; an experiment is not going well and limited time is available to perform the experiment again. The outcome of each experiment is a Gaussian random variable X having a mean equal to the value being sought and variance σ^2. The student decides to

average his experimental outcomes.

$$Y_l = \frac{1}{l}\sum_{k=1}^{l} X_k \ , l = 1,2,\ldots$$

Each outcome X_i is uncorrelated with all other outcomes X_j, $j \neq i$.

(a) Find the mean and correlation function of the stochastic sequence Y_l.

(b) Is Y_l stationary? Indicate your reasoning.

(c) How large must n be to ensure that the probability of the relative error $|(Y_l - \mathsf{E}[Y_l])/\sigma|$ being less than 0.1 is 0.95?

2.39 Predicting the Stock Market

The price of a certain stock can fluctuate during the day while the true value is rising or falling. To facilitate financial decisions, a Wall Street broker decides to use stochastic process theory. The price P_t of a stock is described by

$$P_t = Ct + N_t \ , 0 \leq t < 1$$

where C is the constant our knowledgeable broker is seeking and N_t is a stochastic process describing the random fluctuations. N_t is a white, Gaussian process having spectral height $N_0/2$. The broker decides to estimate C according to:

$$\widehat{C} = \int_0^1 P_t g(t)\, dt$$

where the best function $g(t)$ is to be found.

(a) Find the probability density function of the estimate $\widehat{C}$ for any $g(t)$ the broker might choose.

(b) A simple-minded estimate of C is to use simple averaging (*i.e.*, set $g(t)$ = constant). Find the value of this constant which results in $\mathsf{E}[\widehat{C}] = C$. What is the resulting percentage error as expressed by $\sqrt{\mathsf{var}[\widehat{C}]}/|\,\mathsf{E}[\widehat{C}]|$.

(c) Find $g(t)$ which minimizes the percentage error and yields $\mathsf{E}[\widehat{C}] = C$. How much better is this optimum choice than simple averaging?

2.40 Is DC Present?

To determine the presence or absence of a constant voltage measured in the presence of additive, white Gaussian noise (spectral height $N_0/2$), an engineer decide to compute the average $\bar{V}$ of the measured voltage V_t.

$$\bar{V} = \frac{1}{T}\int_0^T V_t\, dt$$

The value of the constant voltage, if present, is V_0. The presence and absence of the voltage are equally likely to occur.

(a) Derive a good method by which the engineer can use the average to determine the presence or absence of the constant voltage.

(b) Determine the probability that the voltage is present when the engineer's method announces it is.

(c) The engineer decides to improve the performance of his technique by computing the more complicated quantity V given by

$$V = \int_0^T f(t) V_t\, dt$$

What function $f(t)$ maximizes the probability found in part (b)?

2.41 Cross-Correlation Functions

Let X_t be a stationary, zero-mean random process that serves as the input to three linear, time-invariant filters. The power density spectrum of X_t is $\mathscr{S}_X(f) = N_0/2$. The impulse responses of the filters are

$$h_1(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$h_2(t) = \begin{cases} 2e^{-t} & t \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

$$h_3(t) = \begin{cases} \sqrt{2}\sin 2\pi t & 0 \le t \le 2 \\ 0 & \text{otherwise} \end{cases}$$

The output of filter i is denoted by $Y_i(t)$.

(a) Compute $\mathsf{E}[Y_i(t)]$ and $\mathsf{E}[Y_i^2(t)]$ for $i = 1,2,3$.

(b) Compute $R_{XY_2}(t,u)$. Interpret your result.

(c) Is there any pair of processes for which $\mathsf{E}[Y_i(t) \cdot Y_j(t)] = 0$ for all t?

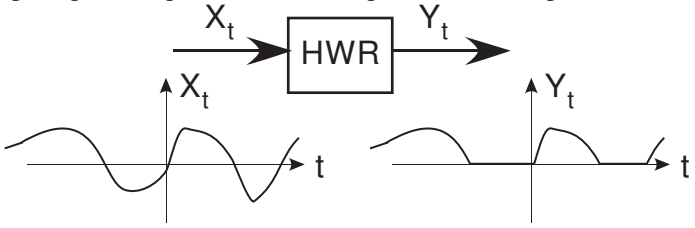
(d) Is there any pair of processes for which $\mathsf{E}[Y_i(t) \cdot Y_j(u)] = 0$ for all t and u?

2.42 Nonlinear Processing

A stationary, zero-mean Gaussian process is passed through a half-wave rectifier, which has an input-output relationship given by

$$Y_t = \begin{cases} X_t & X_t \ge 0 \\ 0 & X_t < 0 \end{cases}$$

In other words, the output equals the positive-valued amplitudes of the input and is zero otherwise.



(a) What is the mean and variance of Y_t? Express your answer in terms of the correlation function of X_t.

(b) Is the output Y_t a stationary process? Indicate why or why not.

(c) What is the cross-correlation function between input and output? Express your answer in terms of the correlation function of X_t.

2.43 An impulse is associated with the occurrence of each event in a stationary Poisson process. This derived process serves as the input to a linear, time-invariant filter having transfer function $H(f)$, which is given by

$$H(f) = 1 - e^{-j\pi fT} + e^{-j2\pi fT}\ , T = \text{constant}\ .$$

(a) What is the mean and covariance function of the input to the filter?

(b) What is the mean and covariance function of the output of the filter?

(c) Now let the filter have any impulse response that has duration T (i.e., $h(t) = 0$, $t < 0$ and $t > T$). Find the impulse response that yields the smallest possible coefficient of variation $v(t)$. The coefficient of variation, defined to be the ratio of the process's standard deviation to its mean at time t, measures the percentage variation of a positive-valued process.

2.44 Linear Vector Spaces and Random Processes

Do the following classes of stochastic processes constitute a linear vector space? If so, indicate the proof; if not, show why not.

1. All stochastic processes.
2. All wide-sense stationary processes.
3. All nonstationary stochastic processes.

2.45 Properties of Inner Products

Show that the inner product of two vectors satisfies the following relationships.

(a) $|\langle x,y\rangle| \leq \|x\| \cdot \|y\|$, the Schwarz inequality.

(b) $\|x+y\| \leq \|x\| + \|y\|$, the triangle inequality.

(c) $\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$, the parallelogram equality.

(d) Find an expression for the inner product $\langle x,y\rangle$ using norms *only*.

2.46 Creating an Inner Product

Let x and y be elements of a normed, linear vector space.

(a) Determine whether the following are valid inner products for the indicated space.

1. $\langle x,y\rangle = \mathbf{x}^t\mathbf{A}\mathbf{y}$ where $\mathbf{A}$ is a nonsingular, $N \times N$ matrix and $\mathbf{x}$, $\mathbf{y}$ are elements of the space of N-dimensional column matrices.
2. $\langle x,y\rangle = \mathbf{x}\mathbf{y}^t$ where $\mathbf{x}$, $\mathbf{y}$ are elements of the space of N-dimensional column matrices.
3. $\langle x,y\rangle = \int_0^T x(t)y(T-t)\,dt$ where x, y are finite-energy signals defined over $[0,T]$.
4. $\langle x,y\rangle = \int_0^T w(t)x(t)y(t)\,dt$ where $w(t)$ is a non-negative function and x, y are finite-energy signals defined over $[0,T]$.
5. $\mathsf{E}[XY]$ where X and Y are real-valued random variables having finite mean-square values.
6. $\mathrm{cov}(X,Y)$, the covariance of the real-valued random variables X and Y. Assume that the random variables have finite mean-square values.

(b) Under what conditions is

$$\int_0^T \int_0^T Q(t,u)x(t)y(u)\,dt\,du$$

a valid inner product for the set of finite-energy functions defined over $[0,T]$?

2.47 Inner Products with Kernels

Let an inner product be defined with respect to the positive-definite, symmetric kernel Q.

$$\langle x,y\rangle_Q = xQy$$

where xQy is the abstract notation for the mapping of the two vectors to a scalar. For example, if $\mathbf{x}$ and $\mathbf{y}$ are column matrices, $\mathbf{Q}$ is a positive-definite square matrix and

$$\langle x,y\rangle_Q = \mathbf{x}^t\mathbf{Q}\mathbf{y}\ .$$

If x and y are defined in L^2, then

$$\langle x,y\rangle_Q = \iint x(t)Q(t,u)y(u)\,dt\,du.$$

Let v denote an eigenvector of Q: $Qv = \lambda v$.

(a) Show that the eigenvectors of a positive-definite, symmetric kernel are orthogonal.

$$\langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0\ , i \neq j\ .$$

(b) Show that these eigenvectors are orthogonal with respect to the inner product generated by Q. Consequently, the eigenvectors are orthogonal with respect to two different inner products.

(c) Let $\widetilde{Q}$ be the inverse kernel associated with Q. If $\mathbf{Q}$ is a matrix, then $\mathbf{Q}\widetilde{\mathbf{Q}} = \mathbf{I}$. If Q is a continuous-time kernel, then

$$\int Q(t,u)\widetilde{Q}(u,v)\,du = \delta(t-v)\ .$$

Show that the eigenvectors of the inverse kernel are equal to those of the kernel. How are the associated eigenvalues of these kernels related to each other?

2.48 Karhunen-Loève Expansions

The purpose of this problem is to derive a general result which describes conditions for an orthonormal basis to result in an uncorrelated representation of a process. Let X denote a stochastic process which has the expansion

$$X = \sum_{i=1}^{\infty} \langle X, \phi_i \rangle \phi_i$$

where $\{\phi_i\}$ denotes a complete, orthonormal basis with respect to the inner product $\langle \cdot, \cdot \rangle$.

$$\langle \phi_i, \phi_j \rangle = \delta_{ij}$$

The process X may be nonstationary and may have non-zero mean.

(a) To require that the representation be an uncorrelated sequence is equivalent to requiring:

$$\mathsf{E}[\langle X, \phi_i \rangle \langle X, \phi_j \rangle] - \mathsf{E}[\langle X, \phi_i \rangle]\,\mathsf{E}[\langle X, \phi_j \rangle] = \lambda_i \delta_{ij}$$

Show that this requirement implies:

$$\mathsf{E}[X\langle X - m_X, \phi_i \rangle] = \lambda_i \phi_i$$

where $m_X = \mathsf{E}[X]$.

(b) Let X be a finite-length stochastic sequence so that it can be considered a random vector. Define the inner product $\langle X, Y \rangle$ to be $\mathbf{X}^t\mathbf{Y}$. Show that above equation is equivalent to

$$\mathbf{K}_X \boldsymbol{\phi} = \lambda \boldsymbol{\phi}\ .$$

(c) Let X be a continuous parameter process so that

$$\langle X, Y \rangle = \int_0^T X_t Y_t\,dt$$

Show that this inner product implies

$$\int_0^T K_X(t,u)\phi(u)\,du = \lambda\,\phi(t).$$

(d) Again let X be a continuous parameter process. However, define the inner product to be

$$\langle X, Y \rangle = \int_0^T \int_0^T Q(t,u) X_t Y_u\,dt\,du.$$

where $Q(t,u)$ is a non-negative definite function. Find the equivalent relationship implied by the requirements of the Karhunen-Love expansion. Under what conditions will the ϕs satisfying this relationship not depend on the covariance function of X?

2.49 Calculating a Karhunen-Loève Expansion

Let the covariance function of a wide-sense stationary process be

$$K_X(\tau) = \begin{cases} 1 - |\tau| & |\tau| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the eigenfunctions and eigenvalues associated with the Karhunen-Love expansion of X_t over $(0, T)$ with $T < 1$.

Chapter 3

Optimization Theory

OPTIMIZATION theory is the study of the *extremal* values of a function: its minima and maxima. Topics in this theory range from conditions for the existence of a unique extremal value to methods—both analytic and numeric—for finding the extremal values and for what values of the independent variables the function attains its extremes. In this book, minimizing an error criterion is an essential step toward deriving optimal signal processing algorithms. An appendix summarizing the key results of optimization theory is essential to understand optimal algorithms.

3.1 Unconstrained Optimization

The simplest optimization problem is to find the minimum of a scalar-valued function of a scalar variable $f(x)$—the so-called *objective function*—and where that minimum is located. Assuming the function is differentiable, the well-known conditions for finding the minima—local and global—are*

$$\frac{df(x)}{dx} = 0$$
$$\frac{d^2f(x)}{dx^2} > 0\,.$$

All values of the independent variable x satisfying these relations are locations of local minima.

Without the second condition, solutions to the first could be either maxima, minima, or inflection points. Solutions to the first equation are termed the *stationary points* of the objective function. To find the *global* minimum—that value (or values) where the function achieves its smallest value—each candidate extremum must be tested: the objective function must be evaluated at each stationary point and the smallest selected. If, however, the objective function can be shown to be strictly convex, then only one solution of $df/dx = 0$ exists and that solution corresponds to the global minimum. The function $f(x)$ is *strictly convex* if, for any choice of x_1, x_2, and the scalar a, $f(ax_1 + (1-a)x_2) < af(x_1) + (1-a)f(x_2)$. Convex objective functions occur often in practice and are more easily minimized because of this property.

When the objective function $f(\cdot)$ depends on a complex variable z, subtleties enter the picture. *If* the function $f(z)$ is differentiable, its extremes can be found in the obvious way: find the derivative, set it equal to zero, and solve for the locations of the extrema. Of particular interest in signal processing are situations where this function is *not* differentiable. In contrast to functions of a real variable, non-differentiable functions of a complex variable occur frequently. The simplest example is $f(z) = |z|^2$. The minimum value of this function obviously occurs at the origin. To calculate this obvious answer, a complication arises: the function $f(z) = z^*$ is not analytic with respect to z and hence not differentiable. More generally, the derivative of a function with respect to a complex-valued variable cannot be evaluated directly when the function depends on the variable's conjugate.

*The maximum of a function is found by finding the minimum of its negative.

This complication can be resolved with either of two methods tailored for optimization problems. The first is to express the objective function in terms of the real and imaginary parts of z and find the function's minimum with respect to these two variables.† This approach is unnecessarily tedious but will yield the solution. The second, more elegant, approach relies on two results from complex variable theory. First, the quantities z and z^* can be treated as independent variables, each considered a constant with respect to the other. A variable and its conjugate are thus viewed as the result of applying an invertible linear transformation to the variable's real and imaginary parts. Thus, if the real and imaginary parts can be considered as independent variables, so can the variable and its conjugate with the advantage that the mathematics is far simpler. In this way, $\partial|z|^2/\partial z = z^*$ and $\partial|z|^2/\partial z^* = z$. Seemingly, the next step to minimizing the objective function is to set the derivatives with respect to each quantity to zero and then solve the resulting pair of equations. As the following theorem suggests, that solution is overly complicated.

Theorem *If the function $f(z,z^*)$ is real-valued and analytic with respect to z and z^*, all stationary points can be found by setting the derivative (in the sense just given) with respect to* either *z or z^* to zero.*

Thus, to find the minimum of $|z|^2$, compute the derivative with respect to either z or z^*. In most cases, the derivative with respect to z^* is the most convenient choice.* Thus, $\partial(|z|^2)/\partial z^* = z$ and the stationary point is $z = 0$. As this objective function is strictly convex, the objective function's sole stationary point is its global minimum.

When the objective function depends on a vector-valued quantity $\mathbf{x}$, the evaluation of the function's stationary points is a simple extension of the scalar-variable case. However, testing stationary points as possible locations for minima is more complicated. The *gradient* of the scalar-valued function $f(\mathbf{x})$ of a vector $\mathbf{x}$ (dimension N) equals an N-dimensional vector where each component is the partial derivative of $f(\cdot)$ with respect to each component of $\mathbf{x}$.

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \text{col}\left[\frac{\partial f(\mathbf{x})}{\partial x_1} \cdots \frac{\partial f(\mathbf{x})}{\partial x_N}\right].$$

For example, the gradient of $\mathbf{x}^t\mathbf{A}\mathbf{x}$ is $\mathbf{A}\mathbf{x} + \mathbf{A}^t\mathbf{x}$. This result is easily derived by expressing the quadratic form as a double sum ($\sum_{ij} A_{ij}x_i x_j$) and evaluating the partials directly. When $\mathbf{A}$ is symmetric, which is often the case, this gradient becomes $2\mathbf{A}\mathbf{x}$.

The gradient "points" in the direction of the maximum rate of increase of the function $f(\cdot)$. This fact is often used in numerical optimization algorithms. The *method of steepest descent* is an iterative algorithm where a candidate minimum is augmented by a quantity proportional to the negative of the objective function's gradient to yield the next candidate.

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \alpha \nabla_{\mathbf{x}} f(\mathbf{x_{k-1}}), \quad \alpha > 0$$

If the objective function is sufficiently "smooth" (there aren't too many minima and maxima), this approach will yield the global minimum. Strictly convex functions are certainly smooth for this method to work.

The gradient of the gradient of $f(\mathbf{x})$, denoted by $\nabla^2_{\mathbf{x}} f(\mathbf{x})$, is a matrix where j^{th} column is the gradient of the j^{th} component of f's gradient. This quantity is known as the *Hessian*, defined to be the matrix of all the second partials of $f(\cdot)$.

$$\left[\nabla^2_{\mathbf{x}} f(\mathbf{x})\right]_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

The Hessian is always a symmetric matrix.

The minima of the objective function $f(\mathbf{x})$ occur when

$$\boxed{\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{0} \quad \text{and} \quad \nabla^2_{\mathbf{x}} f(\mathbf{x}) > 0, \textit{ i.e.}, \text{ Hessian is positive definite.}}$$

Thus, for a stationary point to be a minimum, the Hessian evaluated at that point must be a positive definite matrix. When the objective function is strictly convex, this test need not be performed. For example, the objective function $f(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x}$ is convex whenever $\mathbf{A}$ is positive definite and symmetric.†

†The multi-variate minimization problem is discussed in a few paragraphs.

*Why should this be? In the next few examples, try both and see which you feel is "easier".

†Note that the Hessian of $\mathbf{x}^t\mathbf{A}\mathbf{x}$ is $2\mathbf{A}$.

When the independent vector is complex-valued, the issues discussed in the scalar case also arise. Because of the complex-valued quantities involved, how to evaluate the gradient becomes an issue: is $\nabla_{\mathbf{z}}$ or $\nabla_{\mathbf{z}^*}$ more appropriate? In contrast to the case of complex scalars, the choice in the case of complex vectors is unique.

Theorem *Let $f(\mathbf{z},\mathbf{z}^*)$ be a real-valued function of the vector-valued complex variable $\mathbf{z}$ where the dependence on the variable and its conjugate is explicit. By treating $\mathbf{z}$ and $\mathbf{z}^*$ as independent variables, the quantity pointing in the direction of the maximum rate of change of $f(\mathbf{z},\mathbf{z}^*)$ is $\nabla_{\mathbf{z}^*} f(\mathbf{z})$.*

To show this result, consider the variation of f given by

$$\begin{aligned}\delta f &= \sum_i \left(\frac{\partial f}{\partial z_i}\delta z_i + \frac{\partial f}{\partial z_i^*}\delta z_i^* \right) \\ &= (\nabla_{\mathbf{z}} f)^t\, \delta\mathbf{z} + (\nabla_{\mathbf{z}^*} f)^t\, \delta\mathbf{z}^*\end{aligned}$$

This quantity is concisely expressed as $\delta f = 2\mathrm{Re}\left[(\nabla_{\mathbf{z}^*} f)'\, \delta\mathbf{z}\right]$. By the Schwarz inequality, the maximum value of this variation occurs when $\delta\mathbf{z}$ is in the same direction as $(\nabla_{\mathbf{z}^*} f)$. Thus, the direction corresponding to the largest change in the quantity $f(\mathbf{z},\mathbf{z}^*)$ is in the direction of its gradient with respect to $\mathbf{z}^*$. To implement the method of steepest descent, for example, the gradient with respect to the conjugate *must* be used.

To find the stationary points of a scalar-valued function of a complex-valued vector, we must solve

$$\nabla_{\mathbf{z}^*} f(\mathbf{z}) = \mathbf{0}\,. \tag{3.1}$$

For solutions of this equation to be minima, the Hessian defined to be the matrix of mixed partials given by $\nabla_{\mathbf{z}}(\nabla_{\mathbf{z}^*} f(\mathbf{z}))$ must be positive definite. For example, the required gradient of the objective function $\mathbf{z}'\mathbf{A}\mathbf{z}$ is given by $\mathbf{A}\mathbf{z}$, implying for positive definite $\mathbf{A}$ that a stationary point is $\mathbf{z} = \mathbf{0}$. The Hessian of the objective function is simply $\mathbf{A}$, confirming that the minimum of a quadratic form is always the origin.

3.2 Constrained Optimization

Constrained optimization is the minimization of an objective function subject to constraints on the possible values of the independent variable. Constraints can be either *equality constraints* or *inequality constraints*. Because the scalar-variable case follows easily from the vector one, only the latter is discussed in detail here.

3.2.1 Equality Constraints

The typical constrained optimization problem has the form

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ subject to } \mathbf{g}(\mathbf{x}) = \mathbf{0}\,,$$

where $f(\cdot)$ is the scalar-valued objective function and $\mathbf{g}(\cdot)$ is the vector-valued *constraint function*. Strict convexity of the objective function is *not* sufficient to guarantee a unique minimum; in addition, each component of the constraint must be strictly convex to guarantee that the problem has a unique solution. Because of the constraint, stationary points of $f(\cdot)$ alone may not be solutions to the constrained problem: they may not satisfy the constraints. In fact, solutions to the constrained problem are often *not* stationary points of the objective function. Consequently, the *ad hoc* technique of searching for all stationary points of the objective function that also satisfy the constraint do not work.

The classical approach to solving constrained optimization problems is the method of *Lagrange multipliers*. This approach converts the constrained optimization problem into an unconstrained one, thereby allowing use of the techniques described in the previous section. The *Lagrangian* of a constrained optimization problem is defined to be the scalar-valued function

$$L(\mathbf{x},\boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^t \mathbf{g}(\mathbf{x}).$$

Essentially, the following theorem states that stationary points of the Lagrangian are *potential* solutions of the constrained optimization problem: as always, each candidate solution must be tested to determine which minimizes the objective function.

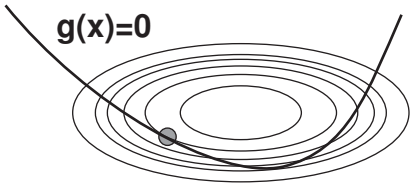


Figure 3.1: The thick line corresponds to the contour of the values of $\mathbf{x}$ satisfying the constraint equation $\mathbf{g}(\mathbf{x}) = \mathbf{0}$. The thinner lines are contours of constant values of the objective function $f(\mathbf{x})$. The contour corresponding to the smallest value of the objective function just tangent to the constraint contour is the solution to the optimization problem with equality constraints.

Theorem *Let $\mathbf{x}_\diamond$ denote a local solution to the constrained optimization problem given above where the gradients $\nabla_{\mathbf{x}} g_1(\mathbf{x}), \ldots, \nabla_{\mathbf{x}} g_M(\mathbf{x})$ of the constraint function's components are linearly independent. There then exists a unique vector $\boldsymbol{\lambda}_\diamond$ such that*

$$\nabla_{\mathbf{x}} L(\mathbf{x}_\diamond, \boldsymbol{\lambda}_\diamond) = \mathbf{0}\,.$$

Furthermore, the quadratic form $\mathbf{y}^t \left[\nabla^2_{\mathbf{x}} L(\mathbf{x}_\diamond, \boldsymbol{\lambda}_\diamond)\right] \mathbf{y}$ is non-negative for all $\mathbf{y}$ satisfying $\left[\nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x})\right]^t \mathbf{y} = 0$.

The latter result in the theorem says that the Hessian of the Lagrangian evaluated at its stationary points is non-negative definite with respect to all vectors "orthogonal" to the gradient of the constraint. This result generalizes the notion of a positive definite Hessian in unconstrained problems.

The rather abstract result of the preceding theorem has a simple geometric interpretation. As shown in Fig. 3.1, the constraint corresponds to a contour in the $\mathbf{x}$ plane. A "contour map" of the objective function indicates those values of $\mathbf{x}$ for which $f(\mathbf{x}) = c$. In this figure, as c becomes smaller, the contours shrink to a small circle in the center of the figure. The solution to the constrained optimization problem occurs when the smallest value of c is chosen for which the contour just touches the constraint contour. At that point, the gradient of the objective function and of the constraint contour are proportional to each other. This proportionality vector is $\boldsymbol{\lambda}_\diamond$, the so-called *Lagrange multiplier*. The Lagrange multiplier's exact value must be such that the constraint is exactly satisfied. Note that the constraint can be tangent to the objective function's contour map for larger values of c. These potential, but erroneous, solutions can be discarded only by evaluating the objective function.

Example

A typical problem arising in signal processing is to minimize $\mathbf{x}^t\mathbf{A}\mathbf{x}$ subject to the linear constraint $\mathbf{c}^t\mathbf{x} = 1$. $\mathbf{A}$ is a positive definite, symmetric matrix (a correlation matrix) in most problems. Clearly, the minimum of the objective function occurs at $\mathbf{x} = \mathbf{0}$, but this cannot satisfy the constraint. The constraint $g(\mathbf{x}) = \mathbf{c}^t\mathbf{x} - 1$ is a scalar-valued one; hence the theorem of Lagrange applies as there are no multiple components in the constraint forcing a check of linear independence. The Lagrangian is

$$L(\mathbf{x}, \lambda) = \mathbf{x}^t\mathbf{A}\mathbf{x} + \lambda\left(\mathbf{c}^t\mathbf{x} - 1\right).$$

Its gradient is $2\mathbf{A}\mathbf{x} + \lambda\mathbf{c}$ with a solution $\mathbf{x}_\diamond = -\lambda_\diamond\mathbf{A}^{-1}\mathbf{c}/2$. To find the value of the Lagrange multiplier, this solution must satisfy the constraint. Imposing the constraint, $\lambda_\diamond\mathbf{c}^t\mathbf{A}^{-1}\mathbf{c} = -2$; thus, $\lambda_\diamond = -2/(\mathbf{c}^t\mathbf{A}^{-1}\mathbf{c})$ and the total solution is

$$\mathbf{x}_\diamond = \frac{\mathbf{A}^{-1}\mathbf{c}}{\mathbf{c}^t\mathbf{A}^{-1}\mathbf{c}}.$$

When the independent variable is complex-valued, the Lagrange multiplier technique can be used *if* care is taken to make the Lagrangian real. If it is not real, we cannot use the theorem {51} that permits computation of stationary points by computing the gradient with respect to $\mathbf{z}^*$ alone. The Lagrangian may not be real-valued even when the constraint is real. Once insured real, the gradient of the Lagrangian with respect to the conjugate of the independent vector can be evaluated and the minimization procedure remains as before.

Example

Consider slight variations to the previous example: let the vector $\mathbf{z}$ be complex so that the objective function is $\mathbf{z}^t\mathbf{A}\mathbf{z}$ where $\mathbf{A}$ is a positive definite, Hermitian matrix and let the constraint be linear, but

vector-valued ($\mathbf{Cz} = \mathbf{c}$). The Lagrangian is formed from the objective function and the real part of the usual constraint term.

$$L(\mathbf{z},\boldsymbol{\lambda}) = \mathbf{z}'\mathbf{A}\mathbf{z} + \boldsymbol{\lambda}'(\mathbf{C}\mathbf{z} - \mathbf{c}) + \boldsymbol{\lambda}^t(\mathbf{C}^*\mathbf{z}^* - \mathbf{c}^*)$$

For the Lagrange multiplier theorem to hold, the gradients of each component of the constraint must be linearly independent. As these gradients are the columns of $\mathbf{C}$, their mutual linear independence means that each constraint vector must not be expressible as a linear combination of the others. We shall assume this portion of the problem statement true. Evaluating the gradient with respect to $\mathbf{z}^*$, keeping $\mathbf{z}$ a constant, and setting the result equal to zero yields

$$\mathbf{A}\mathbf{z}_\diamond + \mathbf{C}'\boldsymbol{\lambda}_\diamond = \mathbf{0}.$$

The solution is $\mathbf{z}_\diamond$ is $-\mathbf{A}^{-1}\mathbf{C}'\boldsymbol{\lambda}_\diamond$. Applying the constraint, we find that $\mathbf{C}\mathbf{A}^{-1}\mathbf{C}'\boldsymbol{\lambda}_\diamond = -\mathbf{c}$. Solving for the Lagrange multiplier and substituting the result into the solution, we find that the solution to the constrained optimization problem is

$$\mathbf{z}_\diamond = \mathbf{A}^{-1}\mathbf{C}'\left(\mathbf{C}\mathbf{A}^{-1}\mathbf{C}'\right)^{-1}\mathbf{c}.$$

The indicated matrix inverses always exist: $\mathbf{A}$ is assumed invertible and $\mathbf{C}\mathbf{A}^{-1}\mathbf{C}'$ is invertible because of the linear independence of the constraints.

3.2.2 Inequality Constraints

When some of the constraints are inequalities, the Lagrange multiplier technique can be used, but the solution must be checked carefully in its details. But first, the optimization problem with equality and inequality constraints is formulated as

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ subject to } \mathbf{g}(\mathbf{x}) = \mathbf{0} \text{ and } \mathbf{h}(\mathbf{x}) \leq \mathbf{0}.$$

As before, $f(\cdot)$ is the scalar-valued objective function and $\mathbf{g}(\cdot)$ is the equality constraint function; $\mathbf{h}(\cdot)$ is the *inequality constraint function*.

The key result which can be used to find the analytic solution to this problem is to first form the Lagrangian in the usual way as $L(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = f(\mathbf{x}) + \boldsymbol{\lambda}^t\mathbf{g}(\mathbf{x}) + \boldsymbol{\mu}^t\mathbf{h}(\mathbf{x})$. The following theorem is the general statement of the Lagrange multiplier technique for constrained optimization problems.

Theorem *Let $\mathbf{x}^\diamond$ be a local minimum for the constrained optimization problem. If the gradients of $\mathbf{g}$'s components and the gradients of those components of $\mathbf{h}(\cdot)$ for which $h_i(\mathbf{x}^\diamond) = 0$ are linearly independent, then*

$$\nabla_{\mathbf{x}} L(\mathbf{x}^\diamond,\boldsymbol{\lambda}^\diamond,\boldsymbol{\mu}^\diamond) = \mathbf{0},$$

where $\mu_i^\diamond \geq 0$ and $\mu_i^\diamond h_i(\mathbf{x}^\diamond) = 0$.

The portion of this result dealing with the inequality constraint differs substantially from that concerned with the equality constraint. Either a component of the constraint equals its maximum value (zero in this case) and the corresponding component of its Lagrange multiplier is non-negative (and is usually positive) *or* a component is less than the constraint and its component of the Lagrange multiplier is zero. This latter result means that some components of the inequality constraint are not as stringent as others and these lax ones do not affect the solution.

The rationale behind this theorem is a technique for converting the inequality constraint into an equality constraint: $h_i(\mathbf{x}) \leq 0$ is equivalent to $h_i(\mathbf{x}) + s_i^2 = 0$. Since the new term, called a *slack variable*, is non-negative, the constraint must be non-positive. With the inclusion of slack variables, the equality constraint theorem can be used and the above theorem results. To prove the theorem, not only does the gradient with respect to $\mathbf{x}$ need to be considered, but also with respect to the vector $\mathbf{s}$ of slack variables. The i^{th} component of the gradient of the Lagrangian with respect to $\mathbf{s}$ at the stationary point is $2\mu_i^\diamond s_i^\diamond = 0$. If in solving the optimization problem, $s_i^\diamond = 0$, the inequality constraint was in reality an equality constraint and that component of the constraint behaves accordingly. As $s_i = [-h_i(\mathbf{x})]^{1/2}$, $s_i = 0$ implies that that component of the inequality constraint must equal zero. On the other hand, if $s_i \neq 0$, the corresponding Lagrange multiplier must be zero.

Example

Consider the problem of minimizing a quadratic form subject to a linear equality constraint and an inequality constraint on the norm of the linear constraint vector's variation.

$$\min_{\mathbf{x}} \mathbf{x}^t\mathbf{A}\mathbf{x} \text{ subject to } (\mathbf{c}+\boldsymbol{\delta})^t\mathbf{x} = 1 \text{ and } \|\boldsymbol{\delta}\|^2 \leq \varepsilon$$

This kind of problem arises in robust estimation. One seeks a solution where one of the "knowns" of the problem, $\mathbf{c}$ in this case, is, in reality, only approximately specified. The independent variables are $\mathbf{x}$ and $\boldsymbol{\delta}$. The Lagrangian for this problem is

$$L(\{\mathbf{x},\boldsymbol{\delta}\},\lambda,\mu) = \mathbf{x}^t\mathbf{A}\mathbf{x} + \lambda\left[(\mathbf{c}+\boldsymbol{\delta})^t\mathbf{x} - 1\right] + \mu\left[\|\boldsymbol{\delta}\|^2 - \varepsilon\right].$$

Evaluating the gradients with respect to the independent variables yields

$$2\mathbf{A}\mathbf{x}^\diamond + \lambda^\diamond(\mathbf{c}+\boldsymbol{\delta}^\diamond) = \mathbf{0}$$
$$\lambda^\diamond\mathbf{x}^\diamond + 2\mu^\diamond\boldsymbol{\delta}^\diamond = \mathbf{0}$$

The latter equation is key. Recall that either $\mu^\diamond = 0$ or the inequality constraint is satisfied with equality. If $\mu^\diamond$ is zero, that implies that $\mathbf{x}^\diamond$ must be zero which will not allow the equality constraint to be satisfied. The inescapable conclusion is that $\|\boldsymbol{\delta}^\diamond\|^2 = \varepsilon$ and that $\boldsymbol{\delta}^\diamond$ is parallel to $\mathbf{x}^\diamond$: $\boldsymbol{\delta}^\diamond = -(\lambda^\diamond/2\mu^\diamond)\mathbf{x}^\diamond$. Using the first equation, $\mathbf{x}^\diamond$ is found to be

$$\mathbf{x}^\diamond = -\frac{\lambda^\diamond}{2}\left(\mathbf{A} - \frac{\lambda^{\diamond 2}}{4\mu^\diamond}\mathbf{I}\right)^{-1}\mathbf{c}.$$

Imposing the constraints on this solution results in a pair of equations for the Lagrange multipliers.

$$\left(\frac{\lambda^{\diamond 2}}{4\mu^\diamond}\right)^2 \mathbf{c}^t\left[\mathbf{A} - \frac{\lambda^{\diamond 2}}{4\mu^\diamond}\mathbf{I}\right]^{-2}\mathbf{c} = \varepsilon$$
$$\mathbf{c}^t\left[\mathbf{A} - \frac{\lambda^{\diamond 2}}{4\mu^\diamond}\mathbf{I}\right]^{-1}\mathbf{c} = -\frac{2}{\lambda^\diamond} - \frac{4\varepsilon\mu^\diamond}{\lambda^{\diamond 2}}$$

Multiple solutions are possible and each must be checked. The rather complicated completion of this example is left to the (numerically oriented) reader.

Problems

3.1 Minumum Variance Estimators

The *minumum-variance estimator* is an example of how constrained optimization theory can be used to derive a "non-standard" estimation procedure. The observations $\mathbf{X}$ consist of a deterministic signal $\mathbf{s}$ corrupted by noise. The standard framework involves finding the unit-sample response $\mathbf{h}$ of the FIR linear filter that will pass the signal with unity gain while minimizing the power of the filtered observations. Such a filter can be considered "optimal:" constraining the filter's signal gain while minimizing the output power would, in effect, minimize that component of the output due to input's non-signal portion. Cast as a constrained optimization problem,

$$\min_{\mathbf{h}} \mathsf{E}\left[\left(\mathbf{h}'\mathbf{X}\right)^2\right] \text{ subject to } \mathbf{h}'\mathbf{s} = 1$$

(a) Find the optimal filter that solves this problem.

(b) Find the variance (power) of the output when this estimator is used.

(c) This result has applications in spectral estimation when we want to estimate the power spectrum of a stochastic sequence at a particular frequency f. Here the "signal" is $\mathbf{s} = \text{col}\left\{1, e^{j2\pi f}, e^{j2\pi 2f}, \ldots, e^{j2\pi(N-1)f}\right\}$, where $N = \dim(\mathbf{s})$. What is the minimum variance power spectrum estimator? What is the expression for the output power, which in this framework, is the spectral power estimate?

Chapter 4

Estimation Theory

In searching for methods of extracting information from noisy observations, this chapter describes *estimation theory*, which has the goal of *extracting from noise-corrupted observations the values of disturbance parameters (noise variance, for example), signal parameters (amplitude or propagation direction), or signal waveforms*. Estimation theory assumes that the observations contain an information-bearing quantity, thereby tacitly assuming that detection-based preprocessing has been performed (in other words, do I have something in the observations worth estimating?). Conversely, detection theory often requires estimation of unknown parameters: Signal presence is assumed, parameter estimates are incorporated into the detection statistic, and consistency of observations and assumptions tested. Consequently, detection and estimation theory form a symbiotic relationship, each requiring the other to yield high-quality signal processing algorithms.

In general, estimation involves establishing an error criterion and minimizing it. The minimization relies on a description of how the quantity to be estimated is expressed in observations. This description takes the form of the conditional density $p_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})$ and, if applicable, any prior knowledge of the quantity's probabilistic description. This description and the prior define the problem; unfortunately, the minimum-error estimation procedure depends heavily on the error criterion as well as the problem formalism. Because of the variety of criterion-dependent estimators, arguments frequently rage about which of several optimal estimators is "better." Each procedure is optimum for its assumed error criterion; thus, the argument becomes which error criterion best describes some intuitive notion of quality. When more *ad hoc*, noncriterion-based procedures* are used, we cannot assess the quality of the resulting estimator relative to the best achievable. As shown later, bounds on the estimation error do exist, but their tightness and applicability to a given situation are always issues in assessing estimator quality. Inventiveness coupled with an understanding of the problem (what types of errors are critically important, for example) are key elements to deciding which estimation procedure "fits" a given problem well.

4.1 Terminology in Estimation Theory

More so than detection theory, estimation theory relies on jargon to characterize the properties of estimators. Without knowing any estimation technique, let's use parameter estimation as our discussion prototype. The parameter estimation problem is to determine from a set of L observations, represented by the L-dimensional vector $\mathbf{X}$, the values of parameters denoted by the vector $\boldsymbol{\theta}$. We write the *estimate* of this parameter vector as $\widehat{\boldsymbol{\theta}}(\mathbf{X})$, where the "hat" denotes the estimate, and the functional dependence on $\mathbf{X}$ explicitly denotes the dependence of the estimate on the observations. This dependence is always present,† but we frequently denote the estimate compactly as $\widehat{\boldsymbol{\theta}}$. Because of the probabilistic nature of the problems considered in this chapter, a parameter estimate is itself a random vector, having its own statistical characteristics. The *estimation error*

*This governmentese phrase concisely means guessing.

†Estimating the value of a parameter given no data may be an interesting problem in clairvoyance, but not in estimation theory.

$\boldsymbol{\varepsilon}(\mathbf{X})$ equals the estimate minus the actual parameter value: $\boldsymbol{\varepsilon}(\mathbf{X}) = \widehat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}$. It too is a random quantity and is often used in the criterion function. For example, the *mean-squared error* is given by $\mathsf{E}[\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}]$; the minimum mean-squared error estimate would minimize this quantity. The mean-squared error matrix is $\mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t]$; on the main diagonal, its entries are the mean-squared estimation errors for each component of the parameter vector, whereas the off-diagonal terms express the correlation between the errors. The mean-squared estimation error $\mathsf{E}[\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}]$ equals the trace of the mean-squared error matrix $\mathrm{tr}\{\mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t]\}$.

Bias. An estimate is said to be *unbiased* if the expected value of the estimate equals the true value of the parameter: $\mathsf{E}[\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}] = \boldsymbol{\theta}$. Otherwise, the estimate is said to be *biased*: $\mathsf{E}[\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}] \neq \boldsymbol{\theta}$. The *bias* $\mathbf{b}(\boldsymbol{\theta})$ is usually considered to be additive, so that $\mathbf{b}(\boldsymbol{\theta}) = \mathsf{E}[\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}] - \boldsymbol{\theta}$. When we have a biased estimate, the bias usually depends on the number of observations L. An estimate is said to be *asymptotically unbiased* if the bias tends to zero for large L: $\lim_{L\to\infty}\mathbf{b} = \mathbf{0}$. An estimate's variance equals the mean-squared estimation error *only* if the estimate is unbiased.

An unbiased estimate has a probability distribution where the mean equals the actual value of the parameter. Should the lack of bias be considered a desirable property? If many unbiased estimates are computed from statistically independent sets of observations having the same parameter value, the average of these estimates will be close to this value. This property does *not* mean that the estimate has less error than a biased one; there exist biased estimates whose mean-squared errors are smaller than unbiased ones. In such cases, the biased estimate is usually asymptotically unbiased. Lack of bias is good, but that is just one aspect of how we evaluate estimators.

Consistency. We term an estimate *consistent* if the mean-squared estimation error tends to zero as the number of observations becomes large: $\lim_{L\to\infty}\mathsf{E}[\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}] = 0$. Thus, a consistent estimate must be at least asymptotically unbiased. Unbiased estimates do exist whose errors never diminish as more data are collected: Their variances remain nonzero no matter how much data are available. Inconsistent estimates may provide reasonable estimates when the amount of data is limited, but have the counterintuitive property that the quality of the estimate does not improve as the number of observations increases. Although appropriate in the proper circumstances (smaller mean-squared error than a consistent estimate over a pertinent range of values of L), consistent estimates are usually favored in practice.

Efficiency. As estimators can be derived in a variety of ways, their error characteristics must always be analyzed and compared. In practice, many problems and the estimators derived for them are sufficiently complicated to render analytic studies of the errors difficult, if not impossible. Instead, numerical simulation and comparison with lower bounds on the estimation error are frequently used instead to assess the estimator performance. An *efficient* estimate has a mean-squared error that equals a particular lower bound: the Cramér-Rao bound. If an efficient estimate exists (the Cramér-Rao bound is the greatest lower bound), it is optimum in the mean-squared sense: No other estimate has a smaller mean-squared error (see §4.2.4 {71} for details).

For many problems no efficient estimate exists. In such cases, the Cramér-Rao bound remains a lower bound, but its value is smaller than that achievable by any estimator. How much smaller is usually not known. However, practitioners frequently use the Cramér-Rao bound in comparisons with numerical error calculations. Another issue is the choice of mean-squared error as the estimation criterion; it may not suffice to pointedly assess estimator performance in a particular problem. Nevertheless, every problem is usually subjected to a Cramér-Rao bound computation and the existence of an efficient estimate considered.

4.2 Parameter Estimation

Determining signal parameter values or a probability distribution's parameters are the simplest estimation problems. Their fundamental utility in signal processing is unquestioned. How do we estimate noise power? What is the best estimator of signal amplitude? Examination of useful estimators, and evaluation of their properties and performances constitute a case study of estimation problems. As expected, many of these issues are interrelated and serve to highlight the intricacies that arise in estimation theory.

All parameters of concern here have unknown values; we classify parameter estimation problems according to whether the parameter is stochastic or not. If it is, the parameter has a probability density known as

the *prior density* (one that applies before the data become available). Choosing the prior, as we have said so often, narrows the problem considerably, suggesting that measurement of the parameter's density would yield something like what was assumed! Said another way, if a prior is not chosen from fundamental considerations (such as the physics of the problem) but from *ad hoc* assumptions, the results could tend to resemble the assumptions you placed on the problem. On the other hand, if the density is not known, the parameter is termed "nonrandom," and its values range unrestricted over some interval. The resulting nonrandom-parameter estimation problem differs greatly from the random-parameter problem. We consider first the former problem, letting θ be a scalar parameter having the prior density $p_\theta(\theta)$. The impact of the *a priori* density becomes evident as various error criteria are established, and an "optimum" estimator is derived.

4.2.1 Random Parameters

Here, we assume that the parameter θ has well-defined probability density $p_\theta(\theta)$ and that $p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$, the description of how the observations depend statistically on the unknowns, is given.

Minimum Mean-Squared Error Estimators

In terms of the densities involved in scalar random-parameter problems, the mean-squared error is given by

$$\mathsf{E}[\varepsilon^2] = \iint (\theta - \widehat{\theta})^2 p_{\mathbf{X},\theta}(\mathbf{x},\theta)\, d\mathbf{x}\, d\theta$$

where $p_{\mathbf{X},\theta}(\mathbf{X},\theta)$ is the joint density of the observations and the parameter. To minimize this integral with respect to $\widehat{\theta}$, we rewrite it using the laws of conditional probability as

$$\mathsf{E}[\varepsilon^2] = \int p_{\mathbf{X}}(\mathbf{x}) \left(\int [\theta - \widehat{\theta}(\mathbf{X})]^2 p_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \right) d\mathbf{x} \tag{4.1}$$

The density $p_{\mathbf{X}}(\cdot)$ is nonnegative. To minimize the mean-squared error, we must minimize the inner integral for each value of $\mathbf{X}$ because the integral is weighted by a positive quantity. We focus attention on the inner integral, which is the conditional expected value of the squared estimation error. The condition, a fixed value of $\mathbf{X}$, implies that we seek that constant — $\widehat{\theta}(\mathbf{X})$ — derived from $\mathbf{X}$ that minimizes the second moment of the random parameter θ. A well-known result from probability theory states that the minimum of $\mathsf{E}[(X-c)^2]$ occurs when the constant c equals the expected value of the random variable X (see §2.1.4 {5}). The inner integral and thereby the mean-squared error is minimized by choosing the estimator to be the conditional expected value of the parameter given the observations.

$$\boxed{\widehat{\theta}_{\text{MMSE}}(\mathbf{X}) = \mathsf{E}[\theta|\mathbf{X}]} \tag{4.2}$$

Thus, a parameter's minimum mean-squared error (*MMSE*) estimate is the parameter's *a posteriori* (after the observations have been obtained) expected value.

The associated conditional probability density $p_{\theta|\mathbf{X}}(\theta|\mathbf{X})$ is not often directly stated in a problem definition and must somehow be derived. In many applications, the likelihood function $p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)$ and the *a priori* density of the parameter are a direct consequence of the problem statement. These densities can be used to find the joint density of the observations and the parameter, enabling us to use Bayes' Rule to find the *a posteriori* density *if* we knew the unconditional probability density of the observations.

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X}) = \frac{p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) p_\theta(\theta)}{p_{\mathbf{X}}(\mathbf{X})}$$

This density $p_{\mathbf{X}}(\mathbf{X})$ is often difficult to determine. Be that as it may, to find the *a posteriori* conditional expected value, it need not be known. The numerator entirely expresses the *a posteriori* density's dependence on θ; the denominator only serves as the scaling factor to yield a unit-area quantity. The expected value is the center-of-mass of the probability density and does *not* depend directly on the "weight" of the density, bypassing calculation of the scaling factor. If not, the *MMSE* estimate can be exceedingly difficult to compute.

Example

Let L statistically independent observations be obtained, each of which is expressed by $X(l) = \theta + N(l)$. Each $N(l)$ is a Gaussian random variable having zero mean and variance σ_N^2. Thus, the unknown parameter in this problem is the mean of the observations. Assume it to be a Gaussian random variable *a priori* (mean m_θ and variance σ_θ^2). The likelihood function is easily found to be

$$p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left\{-\frac{1}{2}\left(\frac{X(l)-\theta}{\sigma_N}\right)^2\right\}$$

so that the *a posteriori* density is given by

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X}) = \frac{\dfrac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{\theta-m_\theta}{\sigma_\theta}\right)^2\right\} \displaystyle\prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left\{-\frac{1}{2}\left(\frac{X(l)-\theta}{\sigma_N}\right)^2\right\}}{p_{\mathbf{X}}(\mathbf{X})}$$

In an attempt to find the expected value of this distribution, lump all terms that do not depend *explicitly* on the quantity θ into a proportionality term.

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X}) \propto \exp\left\{-\frac{1}{2}\left[\frac{\sum(X(l)-\theta)^2}{\sigma_N^2} + \frac{(\theta-m_\theta)^2}{\sigma_\theta^2}\right]\right\}$$

After some manipulation, this expression can be written as

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\theta - \sigma^2\left(\frac{m_\theta}{\sigma_\theta^2} + \frac{\sum X(l)}{\sigma_N^2}\right)\right]^2\right\}$$

where σ^2 is a quantity that succinctly expresses the ratio $\sigma_N^2\sigma_\theta^2/(\sigma_N^2 + L\sigma_\theta^2)$. The form of the *a posteriori* density suggests that it too is Gaussian; its mean, and therefore the *MMSE* estimate of θ, is given by

$$\widehat{\theta}_{\text{MMSE}}(\mathbf{X}) = \sigma^2\left(\frac{m_\theta}{\sigma_\theta^2} + \frac{\sum X(l)}{\sigma_N^2}\right)$$

More insight into the nature of this estimate is gained by rewriting it as

$$\widehat{\theta}_{\text{MMSE}}(\mathbf{X}) = \frac{\sigma_N^2/L}{\sigma_\theta^2 + \sigma_N^2/L} m_\theta + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2/L} \cdot \frac{1}{L}\sum_{l=0}^{L-1} X(l)$$

The term σ_N^2/L is the variance of the averaged observations for a given value of θ; it expresses the squared error encountered in estimating the mean by simple averaging. If this error is much greater than the *a priori* variance of θ ($\sigma_N^2/L \gg \sigma_\theta^2$), implying that the observations are noisier than the variation of the parameter, the *MMSE* estimate ignores the observations and tends to yield the *a priori* mean m_θ as its value. If the averaged observations are less variable than the parameter, the second term dominates, and the average of the observations is the estimate's value. This estimate behavior between these extremes is very intuitive. The detailed form of the estimate indicates how the squared error can be minimized by a linear combination of these extreme estimates.

The conditional expected value of the estimate equals

$$\mathsf{E}[\widehat{\theta}_{\text{MMSE}}|\theta] = \frac{\sigma_N^2/L}{\sigma_\theta^2 + \sigma_N^2/L} m_\theta + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2/L}\theta$$

This estimate is biased because its expected value does not equal the value of the sought-after parameter. It is asymptotically unbiased as the squared measurement error σ_N^2/L tends to zero as L becomes large. The consistency of the estimator is determined by investigating the expected value of

the squared error. Note that the variance of the *a posteriori* density is the quantity σ^2; as this quantity does not depend on $\mathbf{X}$, it also equals the unconditional variance. As the number of observations increases, this variance tends to zero. In concert with the estimate being asymptotically unbiased, the expected value of the estimation error thus tends to zero, implying that we have a consistent estimate.

Maximum a Posteriori Estimators

In those cases in which the expected value of the *a posteriori* density cannot be computed, a related but simpler estimate, the maximum *a posteriori* (*MAP*) estimate, can usually be evaluated. The estimate $\widehat{\theta}_{\text{MAP}}(\mathbf{X})$ equals the location of the maximum of the *a posteriori* density. Assuming that this maximum can be found by evaluating the derivative of the *a posteriori* density, the *MAP* estimate is the solution of the equation

$$\boxed{\left.\frac{\partial p_{\theta|\mathbf{X}}(\theta|\mathbf{X})}{\partial\theta}\right|_{\theta=\widehat{\theta}_{\text{MAP}}} = 0}$$

Any scaling of the density by a positive quantity that depends on $\mathbf{X}$ does not change the location of the maximum. Symbolically, $p_{\theta|\mathbf{X}} = p_{\mathbf{X}|\theta}p_\theta/p_{\mathbf{X}}$; the derivative does not involve the denominator, and this term can be ignored. Thus, the only quantities required to compute $\widehat{\theta}_{\text{MAP}}$ are the likelihood function and the parameter's *a priori* density.

Although not apparent in its definition, the *MAP* estimate does satisfy an error criterion. Define a criterion that is zero over a small range of values about $\varepsilon = 0$ and a positive constant outside that range. Minimization of the expected value of this criterion with respect to $\widehat{\theta}$ is accomplished by centering the criterion function at the maximum of the density. The region having the largest area is thus "notched out," and the criterion is minimized. Whenever the *a posteriori* density is symmetric and unimodal, the *MAP* and *MMSE* estimates coincide. In Gaussian problems, such as the last example, this equivalence is always valid. In more general circumstances, they differ.

Example

Let the observations have the same form as the previous example, but with the modification that the parameter is now uniformly distributed over the interval $[\theta_1, \theta_2]$. The *a posteriori* mean cannot be computed in closed form. To obtain the *MAP* estimate, we need to find the location of the maximum of

$$p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)p_\theta(\theta) = \frac{1}{\theta_2-\theta_1}\prod_{l=0}^{L-1}\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left\{-\frac{1}{2}\left(\frac{X(l)-\theta}{\sigma_N}\right)^2\right\}, \quad \theta_1 \le \theta \le \theta_2$$

Evaluating the logarithm of this quantity does not change the location of the maximum and simplifies the manipulations in many problems. Here, the logarithm is

$$\ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)p_\theta(\theta) = -\ln(\theta_2-\theta_1) - \sum_{l=0}^{L-1}\left(\frac{X(l)-\theta}{\sigma_N}\right)^2 + \ln C, \quad \theta_1 \le \theta \le \theta_2$$

where C is a constant with respect to θ. Assuming that the maximum is interior to the domain of the parameter, the *MAP* estimate is found to be the sample average $\sum X(l)/L$. If the average lies outside this interval, the corresponding endpoint of the interval is the location of the maximum. To summarize,

$$\widehat{\theta}_{\text{MAP}}(\mathbf{X}) = \begin{cases} \theta_1, & \sum_l X(l)/L < \theta_1 \\ \sum_l X(l)/L, & \theta_1 \le \sum_l X(l)/L \le \theta_2 \\ \theta_2, & \theta_2 < \sum_l X(l)/L \end{cases}$$

The *a posteriori* density is not symmetric because of the finite domain of θ. Thus, the *MAP* estimate is not equivalent to the *MMSE* estimate, and the accompanying increase in the mean-squared error is

difficult to compute. When the sample average is the estimate, the estimate is unbiased; otherwise it is biased. Asymptotically, the variance of the average tends to zero, with the consequences that the estimate is unbiased and consistent.

Linear Estimators

We derived the minimum mean-squared error estimator in the previous section with no constraint on the form of the estimator. Depending on the problem, the computations could be a linear function of the observations (which is always the case in Gaussian problems) or nonlinear. Deriving this estimator is often difficult, which limits its application. We consider here a variation of *MMSE* estimation by constraining the estimator to be linear while minimizing the mean-squared estimation error. Such *linear estimators* may not be optimum; the conditional expected value may be nonlinear and it *always* has the smallest mean-squared error. Despite this occasional performance deficit, linear estimators have well-understood properties, they interact well with other signal processing algorithms because of linearity, and they can always be derived, no matter what the problem.

Let the parameter estimate $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ be expressed as $\mathscr{L}[\mathbf{X}]$, where $\mathscr{L}[\cdot]$ is a linear operator: $\mathscr{L}[a_1\mathbf{X}_1 + a_2\mathbf{X}_2] = a_1\mathscr{L}[\mathbf{X}_1] + a_2\mathscr{L}[\mathbf{X}_2]$, a_1, a_2 scalars. Although all estimators of this form are obviously linear, the term *linear estimator* denotes that member of this family that minimizes the mean-squared estimation error.

$$\underset{\widehat{\boldsymbol{\theta}} \in \mathscr{L}[\mathbf{X}]}{\arg\min}\, \mathsf{E}[\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}] = \widehat{\boldsymbol{\theta}}_{\text{LIN}}(\mathbf{X})$$

Because of the transformation's linearity, the theory of linear vector spaces can be fruitfully used to derive the estimator and to specify its properties. One result of that theoretical framework is the well-known *Orthogonality Principle* [78: 407–14]: The linear estimator is that particular linear transformation that yields an estimation error orthogonal to all linear transformations of the data. The orthogonality of the error to *all* linear transformations is termed the "universality constraint." This principle provides us not only with a formal definition of the linear estimator but also with the mechanism to derive it. To demonstrate this intriguing result, let $\langle\cdot,\cdot\rangle$ denote the abstract inner product between two vectors and $\|\cdot\|$ the associated norm.

$$\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$$

For example, if $\mathbf{x}$ and $\mathbf{y}$ are each column matrices having only one column,* their inner product might be defined as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^t\mathbf{y}$. Thus, the linear estimator as defined by the Orthogonality Principle must satisfy

$$\boxed{\mathsf{E}\left[\langle \widehat{\boldsymbol{\theta}}_{\text{LIN}}(\mathbf{X}) - \boldsymbol{\theta}, \mathscr{L}[\mathbf{X}] \rangle\right] = 0, \quad \text{for all linear transformations } \mathscr{L}[\cdot]} \tag{4.3}$$

To see that this principle produces the *MMSE* linear estimator, we express the mean-squared estimation error $\mathsf{E}[\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}] = \mathsf{E}[\|\boldsymbol{\varepsilon}\|^2]$ for *any* choice of linear estimator $\widehat{\boldsymbol{\theta}}$ as

$$\begin{aligned}\mathsf{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2] &= \mathsf{E}[\|(\widehat{\boldsymbol{\theta}}_{\text{LIN}} - \boldsymbol{\theta}) - (\widehat{\boldsymbol{\theta}}_{\text{LIN}} - \widehat{\boldsymbol{\theta}})\|^2] \\ &= \mathsf{E}[\|\widehat{\boldsymbol{\theta}}_{\text{LIN}} - \boldsymbol{\theta}\|^2] + \mathsf{E}[\|\widehat{\boldsymbol{\theta}}_{\text{LIN}} - \widehat{\boldsymbol{\theta}}\|^2] - 2\,\mathsf{E}[\langle \widehat{\boldsymbol{\theta}}_{\text{LIN}} - \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\text{LIN}} - \widehat{\boldsymbol{\theta}} \rangle]\end{aligned}$$

As $\widehat{\boldsymbol{\theta}}_{\text{LIN}} - \widehat{\boldsymbol{\theta}}$ is the difference of two linear transformations, it too is linear and is orthogonal to the estimation error resulting from $\widehat{\boldsymbol{\theta}}_{\text{LIN}}$. As a result, the last term is zero and the mean-squared estimation error is the sum of two squared norms, each of which is, of course, nonnegative. Only the second norm varies with estimator choice; we minimize the mean-squared estimation error by choosing the estimator $\widehat{\boldsymbol{\theta}}$ to be the estimator $\widehat{\boldsymbol{\theta}}_{\text{LIN}}$, which sets the second term to zero.

*There is a confusion as to what a vector is. "Matrices having one column" are colloquially termed vectors as are the field quantities such as electric and magnetic fields. "Vectors" and their associated inner products are taken to be much more general mathematical objects than these. Hence the prose in this section is rather contorted.

The estimation error for the minimum mean-squared linear estimator can be calculated to some degree without knowledge of the form of the estimator. The mean-squared estimation error is given by

$$\begin{aligned}\mathsf{E}[\|\widehat{\boldsymbol{\theta}}_{\text{LIN}}-\boldsymbol{\theta}\|^2] &= \mathsf{E}[\langle\widehat{\boldsymbol{\theta}}_{\text{LIN}}-\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{\text{LIN}}-\boldsymbol{\theta}\rangle] \\ &= \mathsf{E}[\langle\widehat{\boldsymbol{\theta}}_{\text{LIN}}-\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{\text{LIN}}\rangle]+\mathsf{E}[\langle\widehat{\boldsymbol{\theta}}_{\text{LIN}}-\boldsymbol{\theta},-\boldsymbol{\theta}\rangle]\end{aligned}$$

The first term is zero because of the Orthogonality Principle. Rewriting the second term yields a general expression for the *MMSE* linear estimator's mean-squared error.

$$\boxed{\mathsf{E}[\|\boldsymbol{\varepsilon}\|^2]=\mathsf{E}[\|\boldsymbol{\theta}\|^2]-\mathsf{E}[\langle\widehat{\boldsymbol{\theta}}_{\text{LIN}},\boldsymbol{\theta}\rangle]}$$

This error is the difference of two terms. The first, the mean-squared value of the parameter, represents the largest value that the estimation error can be for any reasonable estimator. That error can be obtained by the estimator that ignores the data and has a value of zero. The second term reduces this maximum error and represents the degree to which the estimate and the parameter agree on the average.

Note that the definition of the minimum mean-squared error *linear* estimator makes no explicit assumptions about the parameter estimation problem being solved. This property makes this kind of estimator attractive in many applications where neither the *a priori* density of the parameter vector nor the density of the observations is known precisely. Linear transformations, however, are homogeneous: A zero-valued input yields a zero output. Thus, the linear estimator is especially pertinent to those problems where the expected value of the parameter is zero. If the expected value is nonzero, the linear estimator would not necessarily yield the best result (see Problem 4.20).

Example

Express the first example {58} in vector notation so that the observation vector is written as

$$\mathbf{X}=\theta\mathbf{1}+\mathbf{N}$$

where the vector $\mathbf{1}$ has the form $\mathbf{1}=\text{col}[1,\ldots,1]$. The expected value of the parameter is zero. The linear estimator has the form $\widehat{\theta}_{\text{LIN}}=\mathbf{LX}$, where $\mathbf{L}$ is a $1\times L$ matrix. The Orthogonality Principle states that the linear estimator satisfies

$$\mathsf{E}[(\mathbf{LX}-\theta)^t\mathbf{MX}]=0,\quad \text{for all } 1\times L \text{ matrices } \mathbf{M}$$

To use the Orthogonality Principle to derive an equation implicitly specifying the linear estimator, the "for all linear transformations" phrase must be interpreted. Usually, the quantity specifying the linear transformation must be removed from the constraining inner product by imposing a very stringent but equivalent condition. In this example, this phrase becomes one about matrices. The elements of the matrix $\mathbf{M}$ can be such that each element of the observation vector multiplies each element of the estimation error. Thus, in this problem the Orthogonality Principle means that the expected value of the matrix consisting of all pairwise products of these elements must be zero.

$$\mathsf{E}[(\mathbf{LX}-\theta)\mathbf{X}^t]=\mathbf{0}$$

Thus, two terms must equal each other: $\mathsf{E}[\mathbf{LXX}^t]=\mathsf{E}[\theta\mathbf{X}^t]$. The second term equals $\mathsf{E}[\theta^2]\mathbf{A}^t$ as the additive noise and the parameter are assumed to be statistically independent quantities. The quantity $\mathsf{E}[\mathbf{XX}^t]$ in the first term is the correlation matrix of the observations, which is given by $\mathbf{AA}^t\,\mathsf{E}[\theta^2]+\mathbf{K}_N$. Here, $\mathbf{K}_N$ is the noise covariance matrix, and $\mathsf{E}[\theta^2]$ is the parameter's variance. The quantity $\mathbf{AA}^t$ is a $L\times L$ matrix with each element equaling 1. The noise vector has independent components; the covariance matrix thus equals $\sigma_N^2\mathbf{I}$. The equation that $\mathbf{L}$ must satisfy is therefore given by

$$[\mathbf{L}_1\cdots\mathbf{L}_L]\cdot\begin{bmatrix}\sigma_N^2+\sigma_\theta^2 & \sigma_\theta^2 & \cdots & \sigma_\theta^2\\ \sigma_\theta^2 & \sigma_N^2+\sigma_\theta^2 & \ddots & \vdots\\ \vdots & \ddots & \ddots & \sigma_\theta^2\\ \sigma_\theta^2 & \cdots & \sigma_\theta^2 & \sigma_N^2+\sigma_\theta^2\end{bmatrix}=\begin{bmatrix}\sigma_\theta^2 & \cdots & \sigma_\theta^2\end{bmatrix}$$

The components of $\mathbf{L}$ are equal and are given by $\mathbf{L}_i = \sigma_\theta^2/(\sigma_N^2 + L\sigma_\theta^2)$. Thus, the minimum mean-squared error linear estimator has the form

$$\widehat{\theta}_{\mathrm{LIN}}(\mathbf{X}) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2/L}\frac{1}{L}\sum_l X(l)$$

Note that this result equals the minimum mean-squared error estimate derived earlier under the condition that $\mathsf{E}[\theta] = 0$. Mean-squared error, linear estimators, and Gaussian problems are intimately related to each other. The linear minimum mean-squared error solution to a problem is optimal if the underlying distributions are Gaussian.

4.2.2 Random Parameter Estimation Bounds

Because of the difficulty of finding the minimum mean-squared error estimate $\mathsf{E}[\theta|\mathbf{x}]$, all too often an *ad hoc* estimator $\widehat{\theta}(\mathbf{x})$ is used. To judge estimator quality, a good approach is to determine how the estimator's mean-squared error compares with that of the optimal estimator. The optimal estimator's mean-squared error equals the expected conditional variance; the mean-squared error of *any* cannot be smaller.

$$\mathsf{E}[\varepsilon^2] \geq \mathsf{E}_{\mathbf{X}}[\mathsf{var}[\theta|\mathbf{x}]]$$

This result follows immediately from equations (4.1) and (4.2). Here, $\mathsf{E}_{\mathbf{X}}[\cdot]$ means the expected value with respect to the probability density governing the observations. This quantity may be even harder to compute than the conditional mean. Consequently, more tractable, and therefore useful, bounds on the mean-squared error are sought.

Weinstein-Weiss Family Most useful bounds on mean-squared estimation error belong to the Weinstein-Weiss family [92]. To derive this family of bounds, consider first the scalar parameter case. The family is generated by choosing the function $f(\mathbf{x},\theta)$ that satisfies

$$\int f(\mathbf{x},\theta)p_{\mathbf{X},\theta}(\mathbf{x},\theta)\,d\theta = 0 \quad \forall \mathbf{x} \tag{4.4}$$

As shown subsequently, several choices for $f(\mathbf{x},\theta)$ satisfy this condition. To show why these functions yield interesting bounds on mean-squared estimation error, multiply this equation by any estimator of interest $\widehat{\theta}(\mathbf{x})$ and integrate with respect to $\mathbf{x}$ to yield

$$\mathsf{E}[\widehat{\theta}(\mathbf{x})f(\mathbf{x},\theta)] = \iint \widehat{\theta}(\mathbf{x})f(\mathbf{x},\theta)p_{\mathbf{X},\theta}(\mathbf{x},\theta)\,d\theta d\mathbf{x} = 0$$

Let $g(\theta)$ be any well-behaved function of the parameter* and subtract $\mathsf{E}[g(\theta)f(\mathbf{x},\theta)]$ from both sides of our equation.

$$-\mathsf{E}\left[\left(g(\theta) - \widehat{\theta}(\mathbf{x})\right)f(\mathbf{x},\theta)\right] = -\mathsf{E}[g(\theta)f(\mathbf{x},\theta)]$$

By evaluating the absolute value of both sides and upper-bounding the left side in a simple way, we obtain

$$\mathsf{E}\left[|g(\theta) - \widehat{\theta}(\mathbf{x})| \cdot |f(\mathbf{x},\theta)|\right] \geq |\mathsf{E}[g(\theta)f(\mathbf{x},\theta)]|$$

We now apply the Holder inequality to the left side.†

$$\left(\mathsf{E}\left[|g(\theta) - \widehat{\theta}(\mathbf{x})|^k\right]\right)^{1/k}\left(\mathsf{E}\left[|f(\mathbf{x},\theta)|^{k/(k-1)}\right]\right)^{(k-1)/k} \geq |\mathsf{E}[g(\theta)f(\mathbf{x},\theta)]|$$

*"Well-behaved" means that the expected value $\mathsf{E}[g(\theta)f(\mathbf{x},\theta)]$ is well-defined.

† The Holder inequality states that for non-negative functions $f_1(\cdot)$, $f_2(\cdot)$ and $p(x)$ a probability density function, the expected value of their product has the following upper bound for any real k not equal to one.

$$\int p(x)f_1(x)f_2(x)\,dx \leq \left[\int p(x)[f_1(x)]^k\,dx\right]^{1/k}\left[\int p(x)[f_2(x)]^{k/(k-1)}\,dx\right]^{(k-1)/k}$$

Raising both sides to the k^{th} power and dividing by $\left(\mathsf{E}\left[|f(\mathbf{x},\theta)|^{k/(k-1)}\right]\right)^{(k-1)}$, we obtain a general form for the Weinstein-Weiss bound family.

$$\mathsf{E}\left[|g(\theta)-\widehat{\theta}(\mathbf{x})|^k\right] \geq \frac{|\mathsf{E}[g(\theta)f(\mathbf{x},\theta)]|^k}{\mathsf{E}^{(k-1)}\left[|f(\mathbf{x},\theta)|^{k/(k-1)}\right]}$$

To yield the bound we seek, choose $g(\theta)=\theta$ and $k=2$.

$$\boxed{\mathsf{E}\left[\left(\theta-\widehat{\theta}(\mathbf{x})\right)^2\right] \geq \frac{\mathsf{E}^2[\theta f(\mathbf{x},\theta)]}{\mathsf{E}[f^2(\mathbf{x},\theta)]}} \tag{4.5}$$

The left side is the mean-squared error; the right side is a lower bound that depends on the choice of $f(\mathbf{x},\theta)$ that, of course, must satisfy (4.4). One choice is $f(\mathbf{x},\theta)=\theta-\mathsf{E}[\theta|\mathbf{x}]$. With this choice, we find that the smallest mean-squared error results from using the conditional mean as the estimator.

$$\mathsf{E}\left[\left(\theta-\widehat{\theta}(\mathbf{x})\right)^2\right] \geq \mathsf{E}\left[\left(\theta-\mathsf{E}[\theta|\mathbf{x}]\right)^2\right]$$

Choices for $f(\mathbf{x},\theta)$ that lead to important bounds are listed in the following table.

Name	$f(\mathbf{x},\theta)$	Bound
Cramér-Rao	$\dfrac{\partial \ln p_{\mathbf{X},\theta}(\mathbf{x},\theta)}{\partial\theta}$	$\dfrac{1}{\mathsf{E}\left[\left(\partial \ln p_{\mathbf{X},\theta}(\mathbf{x},\theta)/\partial\theta\right)^2\right]}$
Bhattacharyya	$\displaystyle\sum_{i=1}^{N} a_i \frac{1}{p_{\mathbf{X},\theta}(\mathbf{x},\theta)}\frac{\partial^i p_{\mathbf{X},\theta}(\mathbf{x},\theta)}{\partial\theta^i}$	$(\mathbf{C}^{-1})_{11}$, $\mathbf{C}_{ij}=\mathsf{E}\left[\dfrac{1}{p^2_{\mathbf{X},\theta}(\mathbf{x},\theta)}\dfrac{\partial^i p_{\mathbf{X},\theta}(\mathbf{x},\theta)}{\partial\theta^i}\dfrac{\partial^j p_{\mathbf{X},\theta}(\mathbf{x},\theta)}{\partial\theta^j}\right]$
Weiss-Weinstein	$\Lambda^s(\mathbf{x};\theta+h,\theta)-\Lambda^{1-s}(\mathbf{x};\theta-h,\theta)$, $\Lambda(\mathbf{x};\theta+h,\theta)=p_{\mathbf{X},\theta}(\mathbf{x},\theta+h)/p_{\mathbf{X},\theta}(\mathbf{x},\theta)$	$\dfrac{h^2\,\mathsf{E}^2\left[\Lambda^s(\mathbf{x};\theta+h,\theta)\right]}{\mathsf{E}\left[\left(\Lambda^s(\mathbf{x};\theta+h,\theta)-\Lambda^{1-s}(\mathbf{x};\theta-h,\theta)\right)^2\right]}$

- The Cramér-Rao bound depends on the derivative of $p_{\mathbf{X},\theta}(\mathbf{x},\theta)$ with respect to θ, which indicates that the bound is sensitive to parameter perturbations, possibly not large parameter errors. If large errors can occur, the Cramér-Rao bound remains a *lower* bound but won't be tight.
- The Bhattacharyya bound is an extension of the Cramér-Rao bound, but can be very difficult to compute. The values of the parameters a_i have been optimized to maximize the lower bound, which results in the given form. Of course, when N, the number of terms in the Bhattacharyya bound, equals one, we have the Cramér-Rao bound.
- The Weiss-Weinstein bound is perhaps the most accurate in the sense that it can be attained. The quantity $\Lambda(\mathbf{x};\theta+h,\theta)$ is a special case of the likelihood ratio, a critical quantity in detection. The bound depends on two parameters—s and h—that can be adjusted to produce the largest possible bound. Frequently, $s=1/2$ is the best value, but the best h is problem dependent.

Another bound that has proved useful is the Ziv-Zakai bound, which is discussed on page 65. It is derived using concepts from detection theory. While frequently difficult to calculate, it can be much tighter than the Cramér-Rao bound, as shown in Fig. 4.1 {67}.

Example

Let's return to our example estimation problem.

$$\mathbf{X} = \theta\mathbf{1} + \mathbf{N}$$

To find the Cramér-Rao bound for this problem, we first write the logarithm of the joint density of the observations and the parameter.

$$\ln p_{\mathbf{X},\theta}(\mathbf{x},\theta) = -\frac{L}{2}\ln 2\pi\sigma_N^2 - \frac{1}{2}\ln 2\pi\sigma_\theta^2 - \frac{\sum(x(l)-\theta)^2}{2\sigma_N^2} - \frac{(\theta - m_\theta)^2}{2\sigma_\theta^2}$$

The derivative of this expression with respect to the unknown parameter equals

$$\frac{\partial \ln p_{\mathbf{X},\theta}(\mathbf{x},\theta)}{\partial\theta} = \frac{\sum(x(l)-\theta)}{\sigma_N^2} - \frac{\theta - m_\theta}{\sigma_\theta^2}$$

Squaring, we need to find the expected value of the result.

$$\left(\frac{\partial \ln p_{\mathbf{X},\theta}(\mathbf{x},\theta)}{\partial\theta}\right)^2 = \frac{\sum_{l,m}(x(l)-\theta)(x(m)-\theta)}{\sigma_N^4} - 2\frac{(\theta-m_\theta)\sum(x(l)-\theta)}{\sigma_N^2\sigma_\theta^2} + \frac{(\theta-m_\theta)^2}{\sigma_\theta^4}$$

$$\mathsf{E}\left[\left(\frac{\partial \ln p_{\mathbf{X},\theta}(\mathbf{x},\theta)}{\partial\theta}\right)^2\right] = \frac{L}{\sigma_N^4} + \frac{1}{\sigma_\theta^2}$$

Consequently, the Cramér-Rao bound for the mean-squared error in this problem precisely equals the minimum mean-squared error yielded by the optimal estimator: $\mathsf{E}[\varepsilon^2] \geq \frac{\sigma_N^2\sigma_\theta^2}{L\sigma_\theta^2+\sigma_N^2}$.

To generalize the Weinstein-Weiss bound family to vector parameters, define $\mathbf{f}(\mathbf{x},\boldsymbol{\theta})$ to be a column vector of N family functions: $\mathbf{f}(\mathbf{x},\boldsymbol{\theta}) = \mathrm{col}[f_1(\mathbf{x},\boldsymbol{\theta}),\ldots,f_N(\mathbf{x},\boldsymbol{\theta})]$ where $N = \dim(\boldsymbol{\theta})$. Similarly, define $\mathbf{g}(\mathbf{x}) = \mathrm{col}[g_1(\mathbf{x}),\ldots,g_N(\mathbf{x})]$ and $\widehat{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{col}[\widehat{\theta}_1(\mathbf{x}),\ldots,\widehat{\theta}_N(\mathbf{x})$. Exploiting the previous derivation, define $f(\mathbf{x},\boldsymbol{\theta}) \equiv \mathbf{a}'\mathbf{f}(\mathbf{x},\boldsymbol{\theta})$, $g(\mathbf{x}) \equiv \mathbf{b}'\mathbf{g}(\mathbf{x})$ and $\widehat{\theta}(\mathbf{x},\boldsymbol{\theta}) \equiv \mathbf{b}'\widehat{\boldsymbol{\theta}}(\mathbf{x})$, where $\mathbf{a}$ and $\mathbf{b}$ are column vectors of arbitrary constants. Following the same approach, we find for $k = 2$

$$\mathbf{b}'\,\mathsf{E}\left[(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}(\mathbf{x}))(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}(\mathbf{x}))'\right]\mathbf{b} \geq \frac{(\mathbf{b}'\mathbf{A}\mathbf{a})^2}{\mathbf{a}'\mathbf{B}\mathbf{a}}$$

The matrices $\mathbf{A}$ and $\mathbf{B}$ consist of elements respectively given by

$$\mathbf{A}_{ij} = \mathsf{E}[\theta_i f_j(\mathbf{x},\boldsymbol{\theta})] \quad \mathbf{B}_{ij} = \mathsf{E}[f_i(\mathbf{x},\boldsymbol{\theta})f_j(\mathbf{x},\boldsymbol{\theta})]$$

Because this result holds for all choices of $\mathbf{a}$, choose this vector to maximize the lower bound: assuming $\mathbf{B}$ is invertible, $\mathbf{a} = \mathbf{B}^{-1}\mathbf{A}'\mathbf{b}$.* With this choice, we have that for all choices of $\mathbf{b}$,

$$\mathbf{b}'\,\mathsf{E}\left[(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}(\mathbf{x}))(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}(\mathbf{x}))'\right]\mathbf{b} \geq \mathbf{b}'\mathbf{A}\mathbf{B}^{-1}\mathbf{A}'\mathbf{b}\,.$$

Consequently, the difference of the kernels must be a non-negative definite matrix. We finally arrive at the result that the mean-squared error matrix dominates $\mathbf{A}\mathbf{B}^{-1}\mathbf{A}'$:

$$\boxed{\mathsf{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \succcurlyeq \mathbf{A}\mathbf{B}^{-1}\mathbf{A}'} \tag{4.6}$$

where $\boldsymbol{\varepsilon} = \widehat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}$, the estimation error. The notation $\mathbf{M}_1 \succcurlyeq \mathbf{M}_2$ means that the matrix $\mathbf{M}_1 - \mathbf{M}_2$ is non-negative definite. This result specializes to (4.5) in the scalar case. Note that in order for one matrix to dominate another, the difference of the diagonal elements must be non-negative.

$$\mathbf{M}_1 \succcurlyeq \mathbf{M}_2 \Longrightarrow \mathbf{M}_1(ii) \geq \mathbf{M}_2(ii)$$

*This maximization results from using the Schwarz inequality and recognizing the presence of a Rayleigh quotient (§B.5, {244}) in the inequality.

The diagonal elements of the mean-squared error matrix $\mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$ are the mean-squared errors of each parameter. Consequently, each parameter estimate's mean-squared error has the lower bound

$$\mathsf{E}[\varepsilon_i^2] \geq [\mathbf{A}\mathbf{B}^{-1}\mathbf{A}']_{ii}$$

The following table summarizes lower bounds on the mean-squared estimation error matrix within the Weinstein-Weiss family.

Name	$f_i(\mathbf{x},\boldsymbol{\theta})$	$[\mathbf{A}\mathbf{B}^{-1}\mathbf{A}']$
Cramér-Rao	$\dfrac{\partial \ln p(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_i}$	$\mathbf{F}^{-1}$, $\mathbf{F}_{ij} = \mathsf{E}\left[\dfrac{\partial \ln p(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_i}\dfrac{\partial \ln p(\mathbf{x},\boldsymbol{\theta})}{\partial \theta_j}\right]$
Weiss-Weinstein	$\Lambda^{1/2}(\mathbf{x};\boldsymbol{\theta}+\mathbf{h}_i,\boldsymbol{\theta}) - \Lambda^{1/2}(\mathbf{x};\boldsymbol{\theta}-\mathbf{h}_i,\boldsymbol{\theta})$	$\mathbf{H}\mathbf{G}^{-1}\mathbf{H}^t$, $\mathbf{H} = [\mathbf{h}_1,\ldots,\mathbf{h}_N]$, $\mathbf{G}_{ij} = \dfrac{\mathsf{E}[f_i(\mathbf{x},\boldsymbol{\theta})f_j(\mathbf{x},\boldsymbol{\theta})]}{\mathsf{E}[f_i(\mathbf{x},\boldsymbol{\theta})]\,\mathsf{E}[f_j(\mathbf{x},\boldsymbol{\theta})]}$

Ziv-Zakai Bound The Ziv-Zakai bound is not in the Weiss-Weinstein family [93, 99]. The derivation relies on results from detection theory [21] and provides an example of detection and estimation theory complementing each other to advantage.

First of all, we re-express the mean-squared error in a general way. For a positive-valued random variable X,

$$\mathsf{E}[X^2] = \int_0^\infty x^2 p_X(x)\,dx \overset{\text{integration by parts}}{=} \int_0^\infty 2x\Pr[X \geq x]\,dx \overset{\text{change of variable}}{=} \frac{1}{2}\int_0^\infty \alpha \Pr\left[X \geq \frac{\alpha}{2}\right] d\alpha$$

Therefore, for scalar-valued parameters, the mean-squared error can be expressed as

$$\mathsf{E}[\varepsilon^2] = \frac{1}{2}\int_0^\infty \alpha \Pr\left[|\widehat{\theta} - \theta| \geq \frac{\alpha}{2}\right] d\alpha$$

The critical quantity is $\Pr\left[\left|\widehat{\theta} - \theta\right| \geq \dfrac{\alpha}{2}\right]$, the probability that the absolute value of the estimation error deviates more than a specified value. This quantity can be expressed as

$$\Pr\left[\left|\widehat{\theta} - \theta\right| \geq \frac{\alpha}{2}\right] = \Pr\left[\widehat{\theta} - \theta \geq \frac{\alpha}{2}\right] + \Pr\left[\widehat{\theta} - \theta \leq -\frac{\alpha}{2}\right]$$

Using the law of conditional densities, these two terms can be written as

$$\Pr\left[\widehat{\theta} - \theta \geq \frac{\alpha}{2}\right] = \int_{-\infty}^{\infty} p_\theta(\phi)\Pr\left[\widehat{\theta} \geq \phi + \frac{\alpha}{2}\middle|\, \theta = \phi\right] d\phi$$

$$\Pr\left[\widehat{\theta} - \theta \leq -\frac{\alpha}{2}\right] = \int_{-\infty}^{\infty} p_\theta(\phi+\alpha)\Pr\left[\widehat{\theta} \leq \phi + \frac{\alpha}{2}\middle|\, \theta = \phi+\alpha\right] d\phi$$

Detection theory enters because of the terms $\Pr\left[\widehat{\theta} \geq \phi + \dfrac{\alpha}{2}\middle|\, \theta = \phi\right]$ and $\Pr\left[\widehat{\theta} \leq \phi + \dfrac{\alpha}{2}\middle|\, \theta = \phi+\alpha\right]$ in these expressions. We consider detection problems in great detail in Chapter 5. For now, all we need to know is that the simplest detection problem amounts to deciding which of two descriptions, what we term *models*, best describes the observations. To derive the Ziv-Zakai bound, the two models are that the actual value of the parameter is ϕ or that it is $\phi+\alpha$. The sub-optimal detection rule is to estimate the parameter and compare it to the threshold $\phi + \dfrac{\alpha}{2}$. If the estimate is less than the threshold, we would say that the parameter value is ϕ; if greater, we say the parameter value is $\phi + \dfrac{\alpha}{2}$. One of the probability terms we seek—$\Pr\left[\widehat{\theta} \geq \phi + \dfrac{\alpha}{2}\middle|\, \theta = \phi\right]$—is the probability that our decision rule makes an error when the parameter value equals ϕ; the second term is the probability of the "other" kind of error. The overall probability of error

$P_e(\phi,\phi+\alpha)$ for this decision weights each of these terms by the relative probabilities of the assumed models being true.

$$P_e(\phi,\phi+\alpha)=\frac{p_\theta(\phi)}{p_\theta(\phi)+p_\theta(\phi+\alpha)}\Pr\left[\widehat{\theta}\geq\phi+\frac{\alpha}{2}\,\middle|\,\theta=\phi\right]+\frac{p_\theta(\phi+\alpha)}{p_\theta(\phi)+p_\theta(\phi+\alpha)}\Pr\left[\widehat{\theta}\leq\phi+\frac{\alpha}{2}\,\middle|\,\theta=\phi+\alpha\right]$$

Consequently,

$$\Pr\left[\left|\widehat{\theta}-\theta\right|\geq\frac{\alpha}{2}\right]=\int_{-\infty}^{\infty}\bigl(p_\theta(\phi)+p_\theta(\phi+\alpha)\bigr)P_e(\phi,\phi+\alpha)\,d\phi$$

and therefore for any estimator $\widehat{\theta}(\mathbf{X})$, its mean-squared error is directly related to the probability of error of a detector that uses the estimator in its decision rule.

$$\widehat{\theta}(\mathbf{X})\underset{\theta=\phi}{\overset{\theta=\phi+\alpha}{\gtrless}}\phi+\frac{\alpha}{2} \tag{4.7a}$$

$$\mathsf{E}[\varepsilon^2]=\frac{1}{2}\int_0^\infty\int_{-\infty}^{\infty}\alpha\bigl(p_\theta(\phi)+p_\theta(\phi+\alpha)\bigr)P_e(\phi,\phi+\alpha)\,d\phi\,d\alpha \tag{4.7b}$$

Since the estimator is arbitrary, the decision rule expressed in (4.7a) is suboptimal: a decision rule should be derived from first principles that minimizes the probability of error. Letting $P_e^{\text{opt}}(\phi,\phi+\alpha)$ denote the smallest possible error probability produced by this detector, we arrive at the Ziv-Zaki bound for mean-squared estimation error.

$$\boxed{\mathsf{E}[\varepsilon^2]\geq\frac{1}{2}\int_0^\infty\int_{-\infty}^{\infty}\alpha\bigl(p_\theta(\phi)+p_\theta(\phi+\alpha)\bigr)P_e^{\text{opt}}(\phi,\phi+\alpha)\,d\phi\,d\alpha} \tag{4.8}$$

Calculating this bound requires deriving the probability of error of the optimal decision rule. As the following example shows, the integral required to calculate the bound can be quite complicated.

Example[93]

The time-delay estimation problem is important in many applications, from sonar and radar to wireless communication. Here, the observations consist of the signal $s(l-\tau)$ (τ is an integer here) and additive noise, $l=0,\ldots L$. The signal waveform is assumed known but the delay τ is unknown. The delay is assumed to range uniformly over the observation interval. Combining this restriction to the hypothesized delays yields bounds on both τ and α: $0\leq\alpha<L-\tau$ and $0\leq\tau<L$.* In many cases, the optimum probability of error $P_e^{\text{opt}}(\phi,\phi+\alpha)$ does not depend on α, the time origin of the observations. This lack of dependence is equivalent to ignoring edge effects and simplifies calculation of the bound. Thus, the Ziv-Zakai bound for time-delay estimation relates the mean-squared estimation error for delay to the probability of error incurred by the optimal detector that is deciding whether a nonzero delay is present or not.

$$\boxed{\begin{aligned}\mathsf{E}[\varepsilon^2]&\geq\frac{1}{L}\int_0^L\phi(L-\phi)P_e^{\text{opt}}(\phi)\,d\phi\\&\geq\frac{L^2}{6}P_e^{\text{opt}}(L)-\int_0^L\left(\frac{\Delta^2}{2}-\frac{\Delta^3}{3L}\right)\frac{dP_e}{d\Delta}\,d\Delta\end{aligned}} \tag{4.9}$$

To apply this bound to time-delay estimates (unbiased or not), the optimum probability of error for the type of noise and the relative delay between the two signals must be determined. Substituting this expression into either integral yields the Ziv-Zakai bound.

The general behavior of this bound at parameter extremes can be evaluated in some cases. Note that the Cramér-Rao bound in this problem approaches infinity as either the noise variance grows or

*Note that the unknown parameter is discrete valued. This fact will be ignored in performing the intergrals required for the Ziv-Zakai bound. This approximation can be greeted more readily than differentiation operation required to calculate the Cramér-Rao bound.

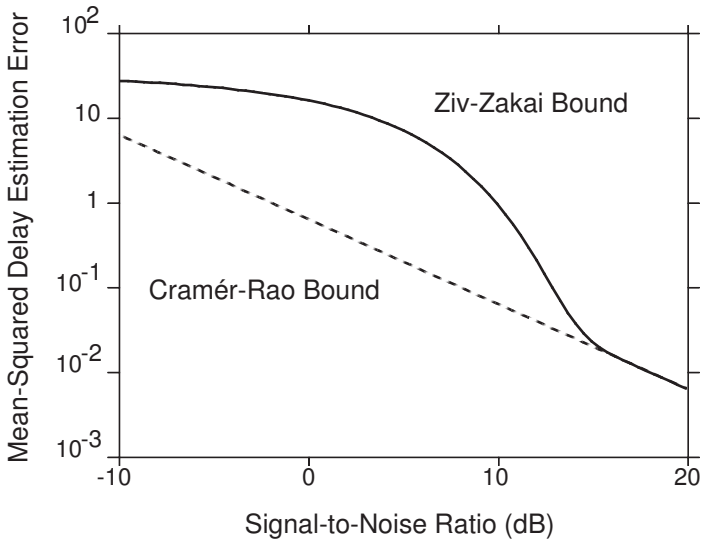


Figure 4.1: The Ziv-Zakai bound and the Cramér-Rao bound for the estimation of the time delay of a signal observed in the presence of Gaussian noise is shown as a function of the signal-to-noise ratio. For this plot, $L = 20$ and $\beta = 2\pi \times 0.2$. The Ziv-Zakai bound is much larger than the Cramér-Rao bound for signal-to-noise ratios less than 13 dB; the Ziv-Zakai bound can be as much as 30 times larger.

the observation interval shrinks to 0 (either forces the signal-to-noise ratio to approach 0). This result is unrealistic as the actual delay is bounded, lying between 0 and L. In this very noisy situation, one should ignore the observations and "guess" *any* reasonable value for the delay; the estimation error is smaller. The probability of error approaches $1/2$ in this situation no matter what the delay Δ may be. Considering the simplified form of the Ziv-Zakai bound, the integral in the second form is 0 in this extreme case.

$$\mathsf{E}[\varepsilon^2] \geq \frac{L^2}{12}$$

The Ziv-Zakai bound is therefore exactly the variance of a random variable uniformly distributed over $[0, L-1]$. The Ziv-Zakai bound thus predicts the size of mean-squared errors more accurately than does the Cramér-Rao bound.

Let the noise be Gaussian of variance σ_N^2 and the signal have energy E. The probability of error resulting from the likelihood ratio test is given by

$$P_e(\Delta) = Q\left(\left[\frac{E}{2\sigma_N^2}\left(1-\rho(\Delta)\right)\right]^{1/2}\right)$$

The quantity $\rho(\Delta)$ is the normalized autocorrelation function of the signal evaluated at the delay Δ.

$$\rho(\Delta) = \frac{1}{E}\sum_l s(l)s(l-\Delta)$$

Evaluation of the Ziv-Zakai bound for a general signal is very difficult in this Gaussian noise case. Fortunately, the normalized autocorrelation function can be bounded by a relatively simple expression to yield a more manageable expression. The key quantity $1-\rho(\Delta)$ in the probability of error expression can be rewritten using Parseval's Theorem.

$$1-\rho(\Delta) = \frac{1}{2\pi E}\int_0^\pi 2|S(\omega)|^2[1-\cos(\omega\Delta)]\,d\omega$$

Using the inequality $1-\cos x \leq x^2/2$, $1-\rho(\Delta)$ is bounded from above by $\min\left\{\Delta^2\beta^2/2, 2\right\}$, where β is the root-mean-squared (*RMS*) signal bandwidth.

$$\beta^2 = \frac{\int_{-\pi}^{\pi} \omega^2 |S(\omega)|^2 \, d\omega}{\int_{-\pi}^{\pi} |S(\omega)|^2 \, d\omega} \tag{4.10}$$

Because $Q(\cdot)$ is a decreasing function, we have $P_e(\Delta) \geq Q\big(\mu \min\{\Delta, \Delta^*\}\big)$, where μ is a combination of all of the constants involved in the argument of $Q(\cdot)$: $\mu = (E\beta^2/4\sigma_N^2)^{1/2}$. This quantity varies with the product of the signal-to-noise ratio E/σ_N^2 and the squared *RMS* bandwidth β^2. The parameter $\Delta^* = 2/\beta$ is known as the critical delay and is twice the reciprocal *RMS* bandwidth. We can use this lower bound for the probability of error in the Ziv-Zakai bound to produce a lower bound on the mean-squared estimation error. The integral in the first form of the bound yields the complicated, but computable result

$$\begin{aligned}\mathsf{E}[\varepsilon^2] \geq{}& \frac{L^2}{6} Q\left(\mu \min\{L, \Delta^*\}\right) + \frac{1}{4\mu^2} P_{\chi_3^2}\left(\mu^2 \min\{L^2, \Delta^{*2}\}\right) \\ &- \frac{2}{3\sqrt{2\pi} L \mu^3} \left[1 - \left(1 + \frac{\mu^2}{2} \min\{L^2, \Delta^{*2}\}\right) \exp\left\{-\mu^2 \min\{L^2, \Delta^{*2}\}/2\right\}\right]\end{aligned}$$

The quantity $P_{\chi_3^2}(\cdot)$ is the probability distribution function of a χ^2 random variable having three degrees of freedom.* Thus, the threshold effects in this expression for the mean-squared estimation error depend on the relation between the critical delay and the signal duration. In most cases, the minimum equals the critical delay Δ^*, with the opposite choice possible for very low bandwidth signals.

The Ziv-Zakai bound and the Cramér-Rao bound for the time-delay estimation problem are shown in Fig. 4.1. Note how the Ziv-Zakai bound matches the Cramér-Rao bound only for large signal-to-noise ratios, where they both equal $1/4\mu^2 = \sigma_N^2/E\beta^2$. For smaller values, the former bound is much larger and provides a better indication of the size of the estimation errors. These errors are because of the "cycle skipping" phenomenon, wherein the estimator is off by multiples of the sinusoidal signal's period. This case occurs when the signal-to-noise ratio drops below some critical value. The Ziv-Zakai bound describes the resulting mean-squared estimation error well, whereas the Cramér-Rao bound only "sees" the errors occurring within the correct period.

4.2.3 Non-Random Parameters: Maximum Likelihood

When the *a priori* density of a parameter is not known or the parameter itself is inconveniently described as a random variable, techniques must be developed that make *no* presumption about the relative possibilities of parameter values. Lacking this knowledge, we can expect the error characteristics of the resulting estimates to be worse than those which can use it.

Since we cannot describe the unknown parameter as random variable — we have no *a priori* density —, we use the notation $p_{\mathbf{X}}(\mathbf{X};\theta)$ for the probability density of the observations. The unknown we seek to estimate is a parameter of the observation's joint density. The maximum likelihood estimate $\widehat{\theta}_{\mathrm{ML}}(\mathbf{X})$ of a non-random parameter is, simply, that parameter value that maximizes the *likelihood function*: $p_{\mathbf{X}}(\mathbf{X};\theta)$. On the surface, $\widehat{\theta}_{\mathrm{ML}}(\mathbf{X})$ does not minimize *any* error criterion.† Assuming that the maximum can be found by evaluating a derivative, $\widehat{\theta}_{\mathrm{ML}}(\mathbf{X})$ is defined by

$$\boxed{\left.\frac{\partial p_{\mathbf{X}}(\mathbf{X};\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}_{\mathrm{ML}}} = 0}.$$

*This distribution function has the "closed-form" expression $P_{\chi_3^2}(x) = 1 - Q(\sqrt{x}) - \sqrt{x/2}\exp\{-x/2\}$.

†More on this point later.

The logarithm of the likelihood function may also be used in this maximization.

Example

Let $X(l)$ be a sequence of independent, identically distributed Gaussian random variables having an unknown mean θ but a known variance σ_N^2. Often, we cannot assign a probability density to a parameter of a random variable's density; we simply do not know what the parameter's value is. Maximum likelihood estimates are often used in such problems. In the specific case here, the derivative of the logarithm of the likelihood function equals

$$\frac{\partial \ln p_{\mathbf{X}}(\mathbf{X};\theta)}{\partial \theta} = \frac{1}{\sigma_N^2}\sum_{l=0}^{L-1}[X(l)-\theta]\,.$$

The solution of this equation is the maximum likelihood estimate, which equals the sample average.

$$\widehat{\theta}_{\mathrm{ML}} = \frac{1}{L}\sum_{l=0}^{L-1}X(l)$$

The expected value of this estimate $\mathsf{E}[\widehat{\theta}_{\mathrm{ML}};\theta]$ equals the actual value θ, showing that the maximum likelihood estimate is unbiased. The mean-square error equals σ_N^2/L and we infer that this estimate is consistent.

Parameter Vectors. The maximum likelihood procedure (as well as the others being discussed) can be easily generalized to situations where more than one parameter must be estimated. Letting $\boldsymbol{\theta}$ denote the parameter vector, the likelihood function is now expressed as $p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})$. The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ of the parameter vector is given by the location of the maximum of the likelihood function (or equivalently of its logarithm). Using derivatives, the calculation of the maximum likelihood estimate becomes

$$\boxed{\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}} = \mathbf{0}},$$

where $\nabla_{\boldsymbol{\theta}}$ denotes the gradient with respect to the parameter vector. This equation means that we must estimate all of the parameters *simultaneously* by setting the partial of the likelihood function with respect to *each* parameter to zero. Given P parameters, we must solve in most cases a set of P nonlinear, simultaneous equations to find the maximum likelihood estimates.

Example

Let's extend the previous example to the situation where neither the mean nor the variance of a sequence of independent Gaussian random variables is known. The likelihood function is, in this case,

$$p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) = \prod_{l=0}^{L-1}\frac{1}{\sqrt{2\pi\theta_2}}\exp\left\{-\frac{1}{2\theta_2}\left[X(l)-\theta_1\right]^2\right\}.$$

Evaluating the partial derivatives of the logarithm of this quantity, we find the following set of two equations to solve for θ_1, representing the mean, and θ_2, representing the variance.*

$$\frac{1}{\theta_2}\sum_{l=0}^{L-1}[X(l)-\theta_1] = 0$$

$$-\frac{L}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{l=0}^{L-1}[X(l)-\theta_1]^2 = 0$$

*The variance rather than the standard deviation is represented by θ_2. The mathematics is messier and the estimator has less attractive properties in the latter case. Problem 4.12 illustrates this point.

The solution of this set of equations is easily found to be

$$\widehat{\theta}_1^{\mathrm{ML}} = \frac{1}{L}\sum_{l=0}^{L-1} X(l)$$

$$\widehat{\theta}_2^{\mathrm{ML}} = \frac{1}{L}\sum_{l=0}^{L-1} \left(X(l) - \widehat{\theta}_1^{\mathrm{ML}}\right)^2$$

The expected value of $\widehat{\theta}_1^{\mathrm{ML}}$ equals the actual value of θ_1; thus, this estimate is unbiased. However, the expected value of the estimate of the variance equals $\theta_2 \cdot (L-1)/L$. The estimate of the variance is biased, but asymptotically unbiased. This bias can be removed by replacing the normalization of L in the averaging computation for $\widehat{\theta}_2^{\mathrm{ML}}$ by $L-1$.

4.2.4 Cramér-Rao Bound

The mean-square estimation error for *any* estimate of a non-random parameter has a lower bound, the *Cramér-Rao bound* [25: pp. 474–477], which defines the ultimate accuracy of *any* estimation procedure. This lower bound, as shown later, is intimately related to the maximum likelihood estimator.

We seek a "bound" on the mean-squared error matrix $\mathbf{M}$ defined to be

$$\mathbf{M} = \mathsf{E}[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^t] = \mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t] .$$

A matrix is "lower bounded" by a second matrix if the difference between the two is a non-negative definite matrix. Define the column matrix $\mathbf{x}$ to be

$$\mathbf{x} = \begin{bmatrix} \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} - \mathbf{b}(\boldsymbol{\theta}) \\ \nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) \end{bmatrix} ,$$

where $\mathbf{b}(\boldsymbol{\theta})$ denotes the column matrix of estimator biases. To derive the Cramér-Rao bound, evaluate $\mathsf{E}[\mathbf{x}\mathbf{x}^t]$.

$$\mathsf{E}[\mathbf{x}\mathbf{x}^t] = \begin{bmatrix} \mathbf{M} - \mathbf{b}\mathbf{b}^t & \mathbf{I} + \nabla_{\boldsymbol{\theta}}^t \mathbf{b} \\ \left(\mathbf{I} + \nabla_{\boldsymbol{\theta}}^t \mathbf{b}\right)^t & \mathbf{F} \end{bmatrix}$$

where $\nabla_{\boldsymbol{\theta}} \mathbf{b}$ represents the matrix of partial derivatives of the bias $[\partial b_i / \partial \theta_j]$. The notation $\nabla_{\boldsymbol{\theta}}^t \mathbf{b}$ represents the matrix $(\nabla_{\boldsymbol{\theta}} \mathbf{b}^t)^t$.

The matrix $\mathbf{F}$ is the *Fisher information matrix*

$$\boxed{\mathbf{F} = \mathsf{E}\left[(\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}))\,(\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}))^t\right]} , \tag{4.11}$$

a *very* important quantity in estimation theory and, surprisingly, many other fields as well. Note that this matrix can alternatively be expressed as

$$\mathbf{F} = -\,\mathsf{E}\left[\nabla_{\boldsymbol{\theta}}\, \nabla_{\boldsymbol{\theta}}^t \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})\right] .$$

The notation $\nabla_{\boldsymbol{\theta}}\, \nabla_{\boldsymbol{\theta}}^t$ means the matrix of all second partials of the quantity it operates on (the gradient of the gradient). This matrix is known as the Hessian. Demonstrating the equivalence of these two forms for the Fisher information is quite easy. Because $\int p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})\, d\mathbf{X} = 1$ for all choices of the parameter vector, the gradient of this expression equals zero. Furthermore, $\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\, p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) / p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})$. Combining these results yields

$$\int (\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}))\, p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})\, d\mathbf{x} = \mathbf{0} .$$

Evaluating the gradient of this quantity (using the chain rule) also yields zero.

$$\int \left[\left(\nabla_{\boldsymbol{\theta}}\, \nabla_{\boldsymbol{\theta}}^t \ln p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})\right) p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) + (\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}))\,(\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}))^t\, p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})\right] d\mathbf{x} = 0$$

or $$\boxed{\mathsf{E}\left[(\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}))\,(\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}))^t\right] = -\,\mathsf{E}\left[\nabla_{\boldsymbol{\theta}}\, \nabla_{\boldsymbol{\theta}}^t \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})\right]}$$

Calculating the expected value for the Hessian form is sometimes easier than finding the expected value of the outer product of the gradient with itself. In the scalar case, we have

$$\boxed{\mathsf{E}\left[\left(\frac{\partial \ln p_{\mathbf{X}}(\mathbf{X};\theta)}{\partial\theta}\right)^2\right] = -\mathsf{E}\left[\frac{\partial^2 \ln p_{\mathbf{X}}(\mathbf{X};\theta)}{\partial\theta^2}\right].}$$

Returning to the derivation of the Cramér-Rao bound, the matrix $\mathsf{E}[\mathbf{x}\mathbf{x}^t]$ is non-negative definite because it is a correlation matrix. Thus, for any column matrix $\boldsymbol{\alpha}$, the quadratic form $\boldsymbol{\alpha}^t\,\mathsf{E}[\mathbf{x}\mathbf{x}^t]\boldsymbol{\alpha}$ is non-negative. Choose a form for $\boldsymbol{\alpha}$ that simplifies the quadratic form. A convenient choice is

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\beta} \\ -\mathbf{F}^{-1}\left(\mathbf{I} + \nabla_{\boldsymbol{\theta}}^t\,\mathbf{b}\right)\boldsymbol{\beta} \end{bmatrix},$$

where $\boldsymbol{\beta}$ is an arbitrary column matrix. The quadratic form becomes in this case

$$\boldsymbol{\alpha}^t\,\mathsf{E}[\mathbf{x}\mathbf{x}^t]\boldsymbol{\alpha} = \boldsymbol{\beta}^t\left[\mathbf{M} - \mathbf{b}\mathbf{b}^t - \left(\mathbf{I} + \nabla_{\boldsymbol{\theta}}\,\mathbf{b}^t\right)^t\mathbf{F}^{-1}\left(\mathbf{I} + \nabla_{\boldsymbol{\theta}}\,\mathbf{b}^t\right)\right]\boldsymbol{\beta}.$$

As this quadratic form must be non-negative, the matrix expression enclosed in brackets must be non-negative definite. We thus obtain the well-known Cramér-Rao bound on the mean-square error matrix.

$$\boxed{\mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t] \succcurlyeq \mathbf{b}(\boldsymbol{\theta})\mathbf{b}^t(\boldsymbol{\theta}) + \left(\mathbf{I} + \nabla_{\boldsymbol{\theta}}\,\mathbf{b}^t\right)^t\mathbf{F}^{-1}\left(\mathbf{I} + \nabla_{\boldsymbol{\theta}}\,\mathbf{b}^t\right)}$$

This form for the Cramér-Rao bound does *not* mean that each term in the matrix of squared errors is greater than the corresponding term in the bounding matrix. As stated earlier, this expression means that the difference between these matrices is non-negative definite. For a matrix to be non-negative definite, each term on the main diagonal must be non-negative. The elements of the main diagonal of $\mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t]$ are the squared errors of the estimate of the individual parameters. Thus, for each parameter, the mean-squared estimation error can be no smaller than

$$\mathsf{E}[(\widehat{\theta}_i - \theta_i)^2] \geq b_i^2(\boldsymbol{\theta}) + \left[\left(\mathbf{I} + \nabla_{\boldsymbol{\theta}}\,\mathbf{b}^t\right)^t\mathbf{F}^{-1}\left(\mathbf{I} + \nabla_{\boldsymbol{\theta}}\,\mathbf{b}^t\right)^t\right]_{ii}.$$

This bound simplifies greatly if the estimator is unbiased ($\mathbf{b} = \mathbf{0}$). In this case, the Cramér-Rao bound becomes

$$\mathsf{E}[(\widehat{\theta}_i - \theta_i)^2] \geq \mathbf{F}_{ii}^{-1}.$$

Thus, the mean-squared error for each parameter in a multiple-parameter, unbiased-estimator problem can be no smaller than the corresponding diagonal term in the *inverse* of the Fisher information matrix. In such problems, the estimate's error characteristics of any parameter become intertwined with the other parameters in a complicated way. Any estimator satisfying the Cramér-Rao bound with equality is said to be *efficient*.

Example

Let's evaluate the Cramér-Rao bound for the example we have been discussing: the estimation of the mean and variance of a length L sequence of statistically independent Gaussian random variables. Let the estimate of the mean θ_1 be the sample average $\widehat{\theta}_1 = \sum X(l)/L$; as shown in the last example, this estimate is unbiased. Let the estimate of the variance θ_2 be the unbiased estimate $\widehat{\theta}_2 = [\sum\left(X(l) - \widehat{\theta}_1\right)^2]/(L-1)$. Each term in the Fisher information matrix $\mathbf{F}$ is given by the expected value of the paired products of derivatives of the logarithm of the likelihood function.

$$\mathbf{F}_{ij} = -\mathsf{E}\left[\frac{\partial^2 \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right]$$

The logarithm of the likelihood function is

$$\ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) = -\frac{L}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}\sum_{l=0}^{L-1}[X(l) - \theta_1]^2,$$

its partial derivatives are

$$\frac{\partial \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_1} = \frac{1}{\theta_2}\sum_{l=0}^{L-1}[X(l)-\theta_1] \tag{4.12}$$

$$\frac{\partial \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_2} = -\frac{L}{2\theta_2}+\frac{1}{2\theta_2^2}\sum_{l=0}^{L-1}[X(l)-\theta_1]^2 \tag{4.13}$$

and its second partials are

$$\frac{\partial^2 \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_1^2} = -\frac{L}{\theta_2} \qquad \frac{\partial^2 \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} = -\frac{1}{\theta_2^2}\sum_{l=0}^{L-1}[X(l)-\theta_1]$$

$$\frac{\partial^2 \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} = -\frac{1}{\theta_2^2}\sum_{l=0}^{L-1}[X(l)-\theta_1] \qquad \frac{\partial^2 \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_2^2} = \frac{L}{2\theta_2^2}-\frac{1}{\theta_2^3}\sum_{l=0}^{L-1}[X(l)-\theta_1]^2$$

The Fisher information matrix has the surprisingly simple form

$$\mathbf{F} = \begin{bmatrix} \frac{L}{\theta_2} & 0 \\ 0 & \frac{L}{2\theta_2^2} \end{bmatrix};$$

its inverse is also a diagonal matrix with the elements on the main diagonal equalling the reciprocal of those in the original matrix. Because of the zero-valued off-diagonal entries in the Fisher information matrix, the errors between the corresponding estimates are not inter-dependent. In this problem, the mean-square estimation errors *using unbiased estimators* can be no smaller than

$$\mathsf{E}[(\widehat{\theta}_1-\theta_1)^2] \geq \frac{\theta_2}{L}$$

$$\mathsf{E}[(\widehat{\theta}_2-\theta_2)^2] \geq \frac{2\theta_2^2}{L}$$

Note that *nowhere* in the preceding example did the form of the estimator enter into the computation of the bound. The only quantity used in the computation of the Cramér-Rao bound is the logarithm of the likelihood function, which is a consequence of the problem statement, not how it is solved. *Only in the case of unbiased estimators is the bound independent of the estimators used.*[*] Because of this property, the Cramér-Rao bound is frequently used to assess the performance limits that can be obtained with an unbiased estimator in a particular problem. When bias is present, the exact form of the estimator's bias explicitly enters the computation of the bound. All too frequently, the unbiased form is used in situations where the *existence* of an unbiased estimator can be questioned. As we shall see, one such problem is time delay estimation, presumably of some importance to the reader. This misapplication of the unbiased Cramér-Rao arises from desperation: the estimator is so complicated and nonlinear that computing the bias is nearly impossible. As shown in Problem 4.14, biased estimators can yield mean-squared errors smaller as well as larger than the unbiased version of the Cramér-Rao bound. Consequently, desperation can yield misinterpretation when a general result is misapplied.

In the single-parameter estimation problem, the Cramér-Rao bound incorporating bias has the well-known form[†]

$$\boxed{\mathsf{E}[\varepsilon^2] \geq b^2 + \frac{\left(1+\frac{db}{d\theta}\right)^2}{-\mathsf{E}\left[\frac{\partial^2 \ln p_{\mathbf{X};\theta}(\mathbf{X};\theta)}{\partial \theta^2}\right]}}.$$

Note that the sign of the bias's derivative determines whether this bound is larger or potentially smaller than the unbiased version, which is obtained by setting the bias term to zero.

[*] That's why we assumed in the example that we used an unbiased estimator for the variance.

[†] Note that this bound differs somewhat from that originally given by Cramér [25: p. 480]; his derivation ignores the additive bias term $\mathbf{b}\mathbf{b}^t$.

Efficiency. An interesting question arises: when, if ever, is the bound satisfied with equality? Recalling the details of the derivation of the bound, equality results when the quantity $\mathsf{E}[\boldsymbol{\alpha}^t \mathbf{x}\mathbf{x}^t \boldsymbol{\alpha}]$ equals zero. As this quantity is the expected value of the square of $\boldsymbol{\alpha}^t \mathbf{x}$, it can only equal zero if $\boldsymbol{\alpha}^t \mathbf{x} = 0$. Substituting in the form of the column matrices $\boldsymbol{\alpha}$ and $\mathbf{x}$, equality in the Cramér-Rao bound results whenever

$$\boxed{\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) = \left[\mathbf{I} + \nabla_{\boldsymbol{\theta}}\, \mathbf{b}^t\right]^{-1} \mathbf{F}[\widehat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta} - \mathbf{b}]}. \tag{4.14}$$

This complicated expression means that only if estimation problems (as expressed by the *a priori* density) have the form of the right side of this equation can the mean-square estimation error equal the Cramér-Rao bound. In particular, the gradient of the log likelihood function can *only* depend on the observations through the estimator. In all other problems, the Cramér-Rao bound is a lower bound but not a tight one; *no* estimator can have error characteristics that equal it. In such cases, we have limited insight into ultimate limitations on estimation error size with the Cramér-Rao bound. However, consider the case where the estimator is unbiased ($\mathbf{b} = \mathbf{0}$). In addition, note the maximum likelihood estimate occurs when the gradient of the logarithm of the likelihood function equals zero: $\nabla_{\boldsymbol{\theta}} \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) = 0$ when $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$. In this case, the condition for equality in the Cramér-Rao bound becomes

$$\mathbf{F}[\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}] = \mathbf{0}\,.$$

As the Fisher information matrix is positive-definite, we conclude that if the estimator equals the maximum likelihood estimator, equality in the Cramér-Rao bound can be achieved. To summarize, if the Cramér-Rao bound *can* be satisfied with equality, *only* the maximum likelihood estimate will achieve it. To use estimation theoretic terminology, *if an efficient estimate exists, it is the maximum likelihood estimate*. This result stresses the importance of maximum likelihood estimates, despite the seemingly *ad hoc* manner by which they are defined.

Example

Consider the Gaussian example being examined so frequently in this section. The components of the gradient of the logarithm of the likelihood function were given earlier by equations (4.12) {72}. These expressions can be rearranged to reveal

$$\begin{bmatrix} \dfrac{\partial \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_1} \\ \dfrac{\partial \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} \dfrac{L}{\theta_2}\left[\left(\dfrac{1}{L}\sum_l X(l)\right) - \theta_1\right] \\ -\dfrac{L}{2\theta_2} + \dfrac{1}{2\theta_2^2}\sum_l [X(l) - \theta_1]^2 \end{bmatrix}.$$

The first component, which corresponds to the estimate of the mean, *is* expressed in the form required for the existence of an efficient estimate. The second component—the partial with respect to the variance θ_2—*cannot* be rewritten in a similar fashion. No unbiased, efficient estimate of the variance exists in this problem. The mean-squared error of the variance's unbiased estimate, but not the maximum likelihood estimate, is lower-bounded by $2\theta_2^2/(L-1)^2$. This error is strictly greater than the Cramér-Rao bound of $2\theta_2^2/L^2$. As no unbiased estimate of the variance can have a mean-squared error equal to the Cramér-Rao bound (no efficient estimate exists for the variance in the Gaussian problem), one presumes that the closeness of the error of our unbiased estimator to the bound implies that it possesses the smallest squared-error of any estimate. This presumption may, of course, be incorrect.

Properties of the Maximum Likelihood Estimator. The maximum likelihood estimate is the most used estimation technique for non-random parameters. Not only because of its close linkage to the Cramér-Rao bound, but also because it has desirable asymptotic properties in the context of *any* Problem [25: pp. 500–6].

1. *The maximum likelihood estimate is at least asymptotically unbiased*. It may be unbiased for any number of observations (as in the estimation of the mean of a sequence of independent random variables) for some problems.

2. *The maximum likelihood estimate is consistent*.
3. *The maximum likelihood estimate is asymptotically efficient*. As more and more data are incorporated into an estimate, the Cramér-Rao bound accurately projects the best attainable error and the maximum likelihood estimate has those optimal characteristics.
4. *Asymptotically, the maximum likelihood estimate is distributed as a Gaussian random variable*. Because of the previous properties, the mean asymptotically equals the parameter and the covariance matrix is $[L\mathbf{F}(\boldsymbol{\theta})]^{-1}$.

Most would agree that a "good" estimator should have these properties. What these results do not provide is an assessment of how many observations are needed for the asymptotic results to apply to some specified degree of precision. Consequently, they should be used with caution; for instance, some other estimator may have a smaller mean-square error than the maximum likelihood for a modest number of observations.

Numerical Methods: EM Algorithm. So far, stress has been placed on analytic solutions for maximum likelihood estimates. Not all problems afford analytic solutions, leaving the only recourse numerical optimization methods. Over the centuries, many numerical methods have been developed; we now understand that each optimization problem is best solved by a problem-dependent set of methods. For example, if the objective function to be maximized is concave, many methods can be used. For example, so-called "hill climbing" methods — ones that depend on the function's gradient — will always find the unique (global) maximum. However, if the objective function is not strictly concave (hilly), such methods will converge to a local maximum, not necessarily the *global* maximum. We encountered such a situation in §3.1 when we have several solutions to setting the gradient to zero: each solution would need to be checked to determine which resulted in the largest value for the objective function. "Hilly" objective functions demand special optimization methods.

To illustrate maximum likelihood problems that don't have analytic solutions, consider a mixture problem. In such problems, each observation is drawn randomly from one of two probability distributions. In a Gaussian framework, let $p^{(1)}(X) = \mathcal{N}(m_1, \sigma_1^2)$ and $p^{(2)}(X) = \mathcal{N}(m_2, \sigma_2^2)$. Letting π be the probability of the first probability density was chosen, the joint pdf of a set of L statistically independent observations is given by

$$p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) = \prod_{l=1}^{L} \left[\pi p^{(1)}(X_l) + (1-\pi) p^{(2)}(X_l) \right] .$$

Here, the parameter vector equals $\boldsymbol{\theta} = [\pi, m_1, \sigma_1^2, m_2, \sigma_2^2]$. Even if the log-likelihood function is considered, setting to zero its gradient with respect to $\boldsymbol{\theta}$ does not yield an analytic solution. Furthermore, the solution will *not* be unique. Note that the mean and variance symbols can be interchanged: which pair is labeled "1" is arbitrary.

If the data drawn from this mixture distribution were labeled (we knew which were drawn from $p^{(1)}(X)$ and $p^{(2)}(X)$), the estimation problem would be easy. We would estimate π by the fraction of observations arising from the pdf we labeled as the first, and the means and variances estimated in the usual way separately for each labeled dataset. Exploiting this insight is the heart of the *expectation-maximization* (EM) algorithm. This algorithm, tailored to maximum likelihood (and MAP as well) estimation, is fairly easy to implement. The good and bad news: the algorithm is guaranteed to increase the likelihood function after each iteration until reaching a stationary point. The downside is that this property means that it is a hill-climbing algorithm and may not find the maximum likelihood estimate in all cases. The idea behind the EM algorithm is to repeatedly find a tight lower bound on the log likelihood function in two phases:

- *E-Step*: Construct the lower bound using Jensen's inequality {23}.
- *M-Step*: Maximize this lower bound.

Let $\mathbf{Z}$ denote a set of latent (hidden) variables that, if observed, would simplify the estimation problem. Assuming statistically independent observed, its log-likelihood function is

$$\sum_l \log p_{X_l}(X_l;\boldsymbol{\theta}) = \sum_l \log \sum_{Z_l} p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta})$$

The second sum runs over all possible values of the latent variables present in the l^{th} observation. Let $Q(Z_l)$ denote an arbitrary probability function for the latent variable Z_l.

$$\begin{aligned}\sum_l \log p_{X_l}(X_l;\boldsymbol{\theta}) &= \sum_l \log \sum_{Z_l} p_{X_l,Z_l;\boldsymbol{\theta}}(X_l,Z_l;\boldsymbol{\theta}) \\ &= \sum_l \log \sum_{Z_l} Q(Z_l)\frac{p_{X_l,Z_l;\boldsymbol{\theta}}(X_l,Z_l;\boldsymbol{\theta})}{Q(Z_l)} \\ &\geq \sum_l \sum_{Z_l} Q(Z_l)\log\frac{p_{X_l,Z_l;\boldsymbol{\theta}}(X_l,Z_l;\boldsymbol{\theta})}{Q(Z_l)} \qquad (*)\end{aligned}$$

The last equation (*) results from applying Jensen's inequality (2.12): $\log(\cdot)$ is a (strictly) concave function and the second equation amounts to an expected value with respect to $Q(Z)$. What we need is to choose $Q(\cdot)$ that maximizes the lower bound, even to the point of achieving equality in the last equation. From the properties of Jensen's inequality, equality only occurs if the "random variable" *with respect to* Z_l is a constant. In our case, this "random variable" is the likelihood function divided by $Q(\cdot)$. Consequently, we require

$$\frac{p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta})}{Q(Z_l)} = c(X_l,\boldsymbol{\theta})$$

This constraint means that $Q(Z_l) \propto p_{X_l,Z_l;\boldsymbol{\theta}}(X_l,Z_l;\boldsymbol{\theta})$. Since $Q(\cdot)$ is a probability function, we find it by normalizing the likelihood function.

$$\begin{aligned}Q(Z_l) &= \frac{p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta})}{\sum_{Z_l} p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta})} \\ &= \frac{p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta})}{p_{X_l}(X_l;\boldsymbol{\theta})} \\ &= p_{Z_l|X_l}(Z_l|X_l;\boldsymbol{\theta})\end{aligned}$$

Consequently, the optimizing choice for $Q(\cdot)$ is the posterior distribution of the latent variables given the (actual) observations and the value of the unknown parameters. Finding this distribution comprises the E-step of the EM algorithm.

Note that we don't know the exact values for the parameters. All that the E-step accomplishes is coming as close as possible to the log-likelihood function based on assumed parameter values. The M-step amounts to finding the maximum likelihood estimate for the unknown parameters from equation (*). We use those estimates to recompute the best lower bound on the log-likelihood, then... Succinctly, with m denoting iteration number,

- *E-Step*: $Q^{(m)}(Z_l) = p_{Z_l|X_l}(Z_l|X_l;\boldsymbol{\theta}^{(m)})$
- *M-Step*: $\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} \sum_l \sum_{Z_l} Q^{(m)}(Z_l)\log\frac{p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta})}{Q^{(m)}(Z_l)}$

What remains is to show that the log-likelihood function increases with each iteration. First of all, note that the likelihood function equals equation (*) when we chose the greatest lower bound for $Q(\cdot)$ when using Jensen's inequality.

$$\log p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}^{(m)}) = \sum_l \sum_{Z_l} Q^{(m)}(Z_l)\log\frac{p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta}^{(m)})}{Q^{(m)}(Z_l)}$$

$\boldsymbol{\theta}(m+1)$ is obtained by maximizing this equation. Consequently,

$$\begin{aligned}\log p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}^{(m+1)}) &\geq \sum_l \sum_{Z_l} Q^{(m)}(Z_l)\log\frac{p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta}^{(m+1)})}{Q^{(m)}(Z_l)} \\ &\geq \sum_l \sum_{Z_l} Q^{(m)}(Z_l)\log\frac{p_{X_l,Z_l}(X_l,Z_l;\boldsymbol{\theta}^{(m)})}{Q^{(m)}(Z_l)} = \log p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}^{(m)})\end{aligned}$$

The first inequality arises from the fact that equation (*) holds for any choice of $\boldsymbol{\theta}$ and $Q(\cdot)$. The second arises because $\boldsymbol{\theta}^{(m+1)}$ is found from maximizing the first equation. Consequently, using any other value for $\boldsymbol{\theta}$ must result in a smaller value.*

4.2.5 Estimation without a Model

When no model is available, how do you develop a parameter estimator? How do you analyze your estimator's characteristics? Such problems arise frequently. As described in §4.8 {128}, one could estimate the probability density function of the data. In many case, the estimated density might appear to be "Gaussian-like," but this observation does guarantee that the data are indeed Gaussian. With or without a pdf estimate, how do you estimate simple quantities, such as the mean?

Most frequently, Gaussian-motivated estimates are used in an *ad hoc* manner. For example, we know that the sample average will converge to the actual mean as the dataset size increases (Law of Large Numbers) except in special circumstances. However, we need to know how well the estimate works with a finite amount of data: is it biased? what is its variance? In many cases, we want to estimate something other than a probability density function's parameter. For example, how do you estimate the Kullback-Leibler distance? Frequently, one uses so-called *plug-in* estimators: if the quantity to be estimated can be written as an expected value, replace the probability function in the expression by the estimated density. In this case, we won't have any idea what the estimator's characteristics are.

Fortunately, recent work in statistics provides a way of estimating the bias and removing it from *any* estimator without requiring additional data. The essence of this procedure, known as the *bootstrap*, is to employ computation as a substitute for a larger dataset. The bootstrap procedure is one of several *resampling* techniques that attempt to provide auxiliary information — variance, bias, and confidence intervals — about a statistical estimate. Another method in this family is the so-called jackknife method, and it has been used for removal of bias in entropy calculations [34]. The book by Efron and Tibshirani [32] provides excellent descriptions of the bootstrap procedure and its theoretical properties.

In a general setting, let $\mathbf{X} = \{X_1, \ldots, X_L\}$ denote a dataset from which we estimate the quantity $\hat{\theta}(\mathbf{X})$. We create a sequence of bootstrap datasets $\mathbf{X}^*_m = \{X^*_{1,m}, \ldots, X^*_{L,m}\}$, $m = 1, \ldots, M$. Each bootstrap dataset has the same number of elements as the original, and is created by selecting elements from the original randomly and with replacement. Thus, elements in the original dataset may or may not appear in a given bootstrap dataset, and each can appear more than once. For example, suppose we had a dataset having four data elements $\{X_1, X_2, X_3, X_4\}$; a possible bootstrap dataset might be $\mathbf{X}^* = \{X_2, X_3, X_1, X_1\}$. The parameter estimated from the m^{th} bootstrap dataset is denoted by $\hat{\theta}^*_m = \hat{\theta}(\mathbf{X}^*_m)$. From the M bootstrap datasets, we estimate the quantity of interest, obtaining the collection of estimates $\{\hat{\theta}^*_1, \ldots, \hat{\theta}^*_M\}$. The suggested number of bootstrap datasets and estimates is a few hundred [32].

The underlying idea behind the bootstrap is the empirical cumulative distribution function.

$$\widehat{P}_X(x) = \frac{1}{L} \sum_l \mathrm{u}(x - X_l)$$

Here, $\mathrm{u}(\cdot)$ is the unit step function. The sum represents the number of data values less than x. As shown in Problem 2.4 {33}, one way of generating a sequence of statistically independent, identically distributed, random variables having a given distribution is to use the cumulative distribution function. The bootstrap estimate is based on generating data according to the data's cumulative distribution. The intent is to generate new datasets having the same probability distribution as the data without knowing the actual distribution. This procedure amounts to creating new datasets by sampling-with-replacement the original data.

The bootstrap estimates *cannot* be used improve the precision of the original estimate, but they can provide estimates of $\theta(\mathbf{X})$'s auxiliary statistics, such as variance, bias, and confidence intervals. First of all, the estimated expected value of the estimator — *not* the expected value of the dataset — is to average the bootstrapped estimators θ^*_m.

$$\bar{\hat{\theta}}^* = \frac{1}{M} \sum_{m=1}^{M} \theta^*_m$$

*This conclusion depends on the log-likelihood function not being too bumpy with many stationary points in close proximity.

The *bootstrap estimate of bias* is found by subtracting from this average the original estimate: bias $= \bar{\hat{\theta}}^* - \theta(\mathbf{X})$. The bootstrap-debiased estimate is, therefore, $2\theta(\mathbf{X}) - \bar{\hat{\theta}}^*$. Calculation of bootstrap-debiased estimates can result in negative estimates when the actual parameter is non-negative. The variance of your estimator—*not* the variance of your dataset—is use the ususal estimate of variance.

$$\widehat{\mathsf{var}}(\widehat{\theta}) = \frac{1}{M}\sum_m (\theta_m^* - \bar{\hat{\theta}}^*)^2$$

Example

A dataset consisting of 20 samples is obtained. The omniscient one knows that they were drawn from an exponential probability density {232} having parameter value equal to one, which means that the theoretical mean and standard deviation both equal to one. Empirically, the sample mean was 0.9673 and the sample standard deviation was 0.8123. Of course, your mortal status suggests you don't know the true values you are estimating. Using 200 bootstrap samples, the bias in your sample mean is estimated to be -0.0053, a small value as you might expect: the sample mean is always an unbiased estimator of the true mean. The estimated bias of the sample standard deviation is somewhat larger 0.0259. However, because of the small sample size, the standard deviation of the mean estimate and standard deviation estimates predicted with the bootstrap are 0.1793 and 0.1352 respectively.

Confidence intervals of level β can also be estimated from the bootstrap estimates by sorting them, and determining which values correspond to the $\beta/2$ and $1-\beta/2$ quantiles. Let $\{\theta^*_{(1)},\ldots,\theta^*_{(M)}\}$ denote the sorted (from smallest to largest) estimates. A raw confidence interval estimate corresponds to $[\theta^*_{(\lfloor M-\beta M/2\rfloor)}, \theta^*_{(\lceil \beta M/2\rceil)}]$. Thus, for the 90% confidence interval, $\beta = 0.9$, and the raw confidence interval corresponds to the 5^{th} and 95^{th} percentiles. Because we want confidence intervals on the bootstrap-debiased estimate rather than the original, we reverse the interval and center it around the debiased estimate: $[2\theta(\mathbf{X}) - \theta^*_{(\lceil \beta M/2\rceil)}, 2\theta(\mathbf{X}) - \theta^*_{(\lfloor M-\beta M/2\rfloor)}]$. According to [32], more bootstrap samples would be needed.

Aside: Shuffling and Sampling with Replacement

You might suspect that a dataset has correlated elements but don't know the precise nature of the correlation. A common approach to decorrelating the dataset is "derive" a new dataset from the original by shuffling. Shuffling—keep every element of the dataset but scrambling the order—amounts to sampling *without* replacement. Mathematically, shuffling amounts to applying a permutation matrix {238} to the dataset. Unfortunately, this approach does not produce a decorrelated sequence. If it did, the correlation matrix of the shuffled dataset should approximate an identity matrix. Letting $\mathbf{P}$ be a permutation matrix and $\mathbf{X}$ a zero-mean dataset, the correlation matrix of the shuffled data will be $\mathsf{E}[\mathbf{PX}(\mathbf{PX})'] = \mathsf{E}[\mathbf{PXX}'\mathbf{P}'] = \mathbf{PK}_X\mathbf{P}'$. You can easily see that this matrix contains all the elements of the data's covariance matrix but shuffled in such a way that the result remains a symmetric matrix. For example, the following permutation matrix when applied to a Toeplitz covariance matrix (the dataset is wide-sense stationary) results in

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{K}_X = \begin{bmatrix} K_0 & K_1 & K_2 & K_3 \\ K_1 & K_0 & K_1 & K_2 \\ K_2 & K_1 & K_0 & K_1 \\ K_3 & K_2 & K_1 & K_0 \end{bmatrix} \quad \mathbf{PK}_X\mathbf{P}' = \begin{bmatrix} K_0 & K_1 & K_1 & K_2 \\ K_1 & K_0 & K_2 & K_1 \\ K_1 & K_2 & K_0 & K_3 \\ K_2 & K_1 & K_3 & K_0 \end{bmatrix}$$

Not only does result not approximate an identity matrix, it can be interpreted as the covariance matrix of a *non-stationary* process (the covariance matrix of the shuffled data is not Toeplitz as the circled super-diagonal shows). All that shuffling a dataset accomplishes is to make it non-stationary.

On the other hand, sampling *with* replacement, as done in performing bootstrap calculations, does result in an uncorrelated data sequence. However, the empirical distribution function that underlies

this approach may not approximate well the true distribution function because of the finite amount of data and statistical dependencies within the dataset.

4.3 Signal Parameter Estimation

One extension of parametric estimation theory necessary for its application to signal processing is the estimation of signal parameters. We assume that we observe a signal $s(l,\boldsymbol{\theta})$, whose characteristics are known save a few parameters $\boldsymbol{\theta}$, in the presence of noise. Signal parameters, such as amplitude, time origin, and frequency if the signal is sinusoidal, must be determined in some way. In many cases of interest, we would find it difficult to justify a particular form for the unknown parameters' *a priori* density. Because of such uncertainties, the minimum mean-squared error and maximum *a posteriori* estimators *cannot* be used in many cases. The minimum mean-squared error *linear* estimator does not require this density, but it is most fruitfully used when the unknown parameter appears in the problem in a linear fashion (such as signal amplitude as we shall see).

4.3.1 Linear Minimum Mean-Squared Error Estimator

The only parameter that is linearly related to a signal is the amplitude. Consider, therefore, the problem where the observations are modeled as

$$X(l) = \theta s(l) + N(l), \quad l = 0,\ldots,L-1$$

The signal waveform $s(l)$ is known and its energy normalized to be unity ($\sum s^2(l) = 1$). The linear estimate of the signal's amplitude is assumed to be of the form $\widehat{\theta} = \sum h(l)X(l)$, where $h(l)$ minimizes the mean-squared error. To use the Orthogonality Principle expressed by Eq. 4.3 {60}, an inner product must be defined for scalars. Little choice avails itself but multiplication as the inner product of two scalars. The Orthogonality Principle states that the estimation error must be orthogonal to all linear transformations defining the kind of estimator being sought.

$$\mathsf{E}\left[\left(\sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l)X(l) - \theta\right)\sum_{k=0}^{L-1} h(k)X(k)\right] = 0 \quad \text{for all } h(\cdot)$$

Manipulating this equation to make the universality constraint more transparent results in

$$\sum_{k=0}^{L-1} h(k)\cdot \mathsf{E}\left[\left(\sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l)X(l) - \theta\right)X(k)\right] = 0 \quad \text{for all } h(\cdot)$$

Written in this way, the expected value must be 0 for each value of k to satisfy the constraint. Thus, the quantity $h_{\mathrm{LIN}}(\cdot)$ of the estimator of the signal's amplitude must satisfy

$$\sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l)\,\mathsf{E}[X(l)X(k)] = \mathsf{E}[\theta X(k)] \quad \text{for all } k$$

Assuming that the signal's amplitude has zero mean and is statistically independent of the zero-mean noise, the expected values in this equation are given by

$$\mathsf{E}[X(l)X(k)] = \sigma_\theta^2 s(l)s(k) + K_N(k,l)$$
$$\mathsf{E}[\theta X(k)] = \sigma_\theta^2 s(k)$$

where $K_N(k,l)$ is the covariance function of the noise. The equation that must be solved for the unit-sample response $h_{\mathrm{LIN}}(\cdot)$ of the optimal linear *MMSE* estimator of signal amplitude becomes

$$\boxed{\sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l)K_N(k,l) = \sigma_\theta^2 s(k)\left[1 - \sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l)s(l)\right] \quad \text{for all } k}$$

This equation is easily solved once phrased in matrix notation. Letting $\mathbf{K}_N$ denote the covariance matrix of the noise, $\mathbf{s}$ the signal vector, and $\mathbf{h}_{\text{LIN}}$ the vector of coefficients, this equation becomes

$$\mathbf{K}_N\mathbf{h}_{\text{LIN}} = \sigma_\theta^2\left(1-\mathbf{s}^t\mathbf{h}_{\text{LIN}}\right)\mathbf{s}$$

or

$$\mathbf{h}_{\text{LIN}} = \sigma_\theta^2\left(1-\mathbf{s}^t\mathbf{h}_{\text{LIN}}\right)\mathbf{K}_N^{-1}\mathbf{s}$$

While this equation does not lead directly to a solution, let's assume that the solution has the form $\mathbf{h}_{\text{LIN}} = c\mathbf{K}_N^{-1}\mathbf{s}$, c is a scalar constant. This proposed solution satisfies the equation; the *MMSE* estimate of signal amplitude is

$$\boxed{\mathbf{h}_{\text{LIN}} = \frac{\sigma_\theta^2}{1+\sigma_\theta^2\mathbf{s}^t\mathbf{K}_N^{-1}\mathbf{s}}\mathbf{K}_N^{-1}\mathbf{s}}.$$

The mean-squared estimation error of signal amplitude is given by

$$\mathsf{E}[\varepsilon^2] = \sigma_\theta^2 - \mathsf{E}\left[\theta\sum_{l=0}^{L-1}h_{\text{LIN}}(l)X(l)\right]$$

Substituting the vector expression for $\mathbf{h}_{\text{LIN}}$ yields the result that the mean-squared estimation error equals the proportionality constant c defined earlier.

$$\mathsf{E}[\varepsilon^2] = \frac{\sigma_\theta^2}{1+\sigma_\theta^2\mathbf{s}^t\mathbf{K}_N^{-1}\mathbf{s}}$$

If we had not assumed the noise to be Gaussian, the estimator would be unchanged. This invariance occurs because the linear *MMSE* estimator requires *no* assumptions on the noise's amplitude characteristics.

Example

Let the noise be white so that its covariance matrix is proportional to the identity matrix ($\mathbf{K}_N = \sigma_N^2\mathbf{I}$). The weighting factor in the minimum mean-squared error linear estimator is proportional to the signal waveform.

$$h_{\text{LIN}}(l) = \frac{\sigma_\theta^2}{\sigma_N^2+\sigma_\theta^2}s(l) \qquad \widehat{\theta}_{\text{LIN}} = \frac{\sigma_\theta^2}{\sigma_N^2+\sigma_\theta^2}\sum_{l=0}^{L-1}s(l)X(l)$$

This proportionality constant depends only on the relative variances of the noise and the parameter. *If* the noise variance can be considered to be much smaller than the *a priori* variance of the amplitude, then this constant does not depend on these variances and equals unity. Otherwise, the variances must be known.

We find the mean-squared estimation error to be

$$\mathsf{E}[\varepsilon^2] = \frac{\sigma_\theta^2}{1+\sigma_\theta^2/\sigma_N^2}$$

This error is significantly reduced from its nominal value σ_θ^2 only when the variance of the noise is small compared with the *a priori* variance of the amplitude. Otherwise, this admittedly optimum amplitude estimate performs poorly, and we might as well as have ignored the data and "guessed" that the amplitude was zero.*

---

*In other words, the problem is difficult in this case.

4.3.2 Maximum Likelihood Estimators

Many situations are either not well suited to linear estimation procedures, or the parameter is not well described as a random variable. For example, signal delay is observed nonlinearly and usually no *a priori* density can be assigned. In such cases, maximum likelihood estimators are more frequently used. Because of the Cramér-Rao bound, fundamental limits on parameter estimation performance can be derived for *any* signal parameter estimation problem where the parameter is not random.

Assume that the data are expressed as a signal observed in the presence of additive Gaussian noise.

$$X(l) = s(l,\boldsymbol{\theta}) + N(l), \quad l = 0,\ldots,L-1$$

The vector of observations $\mathbf{X}$ is formed from the data in the obvious way. Evaluating the logarithm of the observation vector's joint density,

$$\ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) = -\frac{1}{2}\ln\det[2\pi\mathbf{K}_N] - \frac{1}{2}[\mathbf{X}-\mathbf{s}(\boldsymbol{\theta})]^t\mathbf{K}_N^{-1}[\mathbf{X}-\mathbf{s}(\boldsymbol{\theta})]$$

where $\mathbf{s}(\boldsymbol{\theta})$ is the signal vector having P unknown parameters, and $\mathbf{K}_N$ is the covariance matrix of the noise. The partial derivative of this likelihood function with respect to the i^{th} parameter θ_i is

$$\frac{\partial \ln p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial\theta_i} = [\mathbf{X}-\mathbf{s}(\boldsymbol{\theta})]^t\mathbf{K}_N^{-1}\frac{\partial\mathbf{s}(\boldsymbol{\theta})}{\partial\theta_i}\ .$$

If the maximum of the likelihood function can be found by setting its gradient to $\mathbf{0}$, the maximum likelihood estimate of the parameter vector is the solution of the set of equations

$$\boxed{[\mathbf{X}-\mathbf{s}(\boldsymbol{\theta})]^t\mathbf{K}_N^{-1}\frac{\partial\mathbf{s}(\boldsymbol{\theta})}{\partial\theta_i}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}} = 0, \quad i = 1,\ldots,P}$$

The Cramér-Rao bound depends on the evaluation of the Fisher information matrix $\mathbf{F}$. The elements of this matrix are found to be

$$F_{ij} = \frac{\partial\mathbf{s}^t(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{K}_N^{-1}\frac{\partial\mathbf{s}(\boldsymbol{\theta})}{\partial\theta_j}, \quad i,j = 1,\ldots,P \tag{4.15}$$

Further computation of the Cramér-Rao bound's components is problem dependent if more than one parameter is involved and the off-diagonal terms of $\mathbf{F}$ are nonzero. If only one parameter is unknown, the Cramér-Rao bound is given by

$$\boxed{\mathsf{E}[\varepsilon^2] \geq b^2(\theta) + \frac{\left(1+\dfrac{db(\theta)}{d\theta}\right)^2}{\dfrac{\partial\mathbf{s}^t(\theta)}{\partial\theta}\mathbf{K}_N^{-1}\dfrac{\partial\mathbf{s}(\theta)}{\partial\theta}}}$$

When the signal depends on the parameter nonlinearly (which constitute the interesting cases), the maximum likelihood estimate is usually biased. Thus, the numerator of the expression for the bound cannot be ignored. One interesting special case occurs when the noise is white. The Cramér-Rao bound becomes

$$\mathsf{E}[\varepsilon^2] \geq b^2(\theta) + \frac{\sigma_N^2\left(1+\dfrac{db(\theta)}{d\theta}\right)^2}{\displaystyle\sum_{l=0}^{L-1}\left(\frac{\partial s(l,\theta)}{\partial\theta}\right)^2}$$

The derivative of the signal with respect to the parameter can be interpreted as the sensitivity of the signal to the parameter. The mean-squared estimation error depends on the "integrated" squared sensitivity: The greater this sensitivity, the smaller the bound.

For an efficient estimate of a signal parameter to exist, the estimate must satisfy the condition we derived earlier (Eq. 4.14 {73}).

$$[\nabla_{\boldsymbol{\theta}}\mathbf{s}(\boldsymbol{\theta})]^t\mathbf{K}_N^{-1}[\mathbf{X}-\mathbf{s}(\boldsymbol{\theta})] \stackrel{?}{=} \left[\mathbf{I}+\nabla_{\boldsymbol{\theta}}\mathbf{b}^t\right]^{-1}[\nabla_{\boldsymbol{\theta}}\mathbf{s}(\boldsymbol{\theta})]^t\mathbf{K}_N^{-1}[\nabla_{\boldsymbol{\theta}}\mathbf{s}(\boldsymbol{\theta})][\widehat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}-\mathbf{b}]$$

Because of the complexity of this requirement, we quite rightly question the existence of any efficient estimator, especially when the signal depends nonlinearly on the parameter (see Problem 4.21).

Example

Let the unknown parameter be the signal's amplitude; the signal is expressed as $\theta s(l)$ and is observed in the presence of additive noise. The maximum likelihood estimate of the amplitude is the solution of the equation

$$[\mathbf{X} - \widehat{\theta}_{\mathrm{ML}}\mathbf{s}]^t \mathbf{K}_N^{-1}\mathbf{s} = 0$$

The form of this equation suggests that the maximum likelihood estimate is efficient. The amplitude estimate is given by

$$\widehat{\theta}_{\mathrm{ML}} = \frac{\mathbf{X}^t \mathbf{K}_N^{-1}\mathbf{s}}{\mathbf{s}^t \mathbf{K}_N^{-1}\mathbf{s}}$$

The form of this estimator is precisely that of the matched filter derived in the colored-noise situation (see Eq. 5.13 {180}). The expected value of the estimate equals the actual amplitude. Thus the bias is zero and the Cramér-Rao bound is given by

$$\mathsf{E}[\varepsilon^2] \geq \left(\mathbf{s}^t \mathbf{K}_N^{-1}\mathbf{s}\right)^{-1}$$

The condition for an efficient estimate becomes

$$\mathbf{s}^t \mathbf{K}_N^{-1}(\mathbf{X} - \theta\mathbf{s}) \stackrel{?}{=} \mathbf{s}^t \mathbf{K}_N^{-1}\mathbf{s} \cdot (\widehat{\theta}_{\mathrm{ML}} - \theta)$$

whose veracity we can easily verify.

In the special case where the noise is white, the estimator has the form $\widehat{\theta}_{\mathrm{ML}} = \mathbf{X}^t\mathbf{s}$, and the Cramér-Rao bound equals σ_N^2 (the nominal signal is assumed to have unit energy). The maximum likelihood estimate of the amplitude has *fixed* error characteristics that do not depend on the actual signal amplitude. A signal-to-noise ratio for the estimate, defined to be $\theta^2 / \mathsf{E}[\varepsilon^2]$, equals the signal-to-noise ratio of the observed signal.

When the amplitude is well described as a random variable, its linear minimum mean-squared error estimator has the form

$$\widehat{\theta}_{\mathrm{LIN}} = \frac{\sigma_\theta^2 \mathbf{X}^t \mathbf{K}_N^{-1}\mathbf{s}}{1 + \sigma_\theta^2 \mathbf{s}^t \mathbf{K}_N^{-1}\mathbf{s}}$$

which we found in the white-noise case becomes a weighted version of the maximum likelihood estimate (see the example {79}).

$$\widehat{\theta}_{\mathrm{LIN}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2}\mathbf{X}^t\mathbf{s}$$

Seemingly, these two estimators are being used to solve the same problem: Estimating the amplitude of a signal whose waveform is known. They make very different assumptions, however, about the nature of the unknown parameter; in one it is a random variable (and thus it has a variance), whereas in the other it is not (and variance makes no sense). Despite this fundamental difference, the computations for each estimator are equivalent. It is reassuring that different approaches to solving similar problems yield similar procedures.

4.3.3 Time-Delay Estimation

An important signal parameter estimation problem is time-delay estimation. Here the unknown is the time origin of the signal: $s(l, \theta) = s(l - \theta)$. The duration of the signal (the domain over which the signal is defined) is assumed brief compared with the observation interval L. Although in continuous time the signal delay is a continuous-valued variable, in discrete time it is not. Consequently, the maximum likelihood estimate *cannot* be found by differentiation, and we must determine the maximum likelihood estimate of signal delay by the most fundamental expression of the maximization procedure. Assuming Gaussian noise, the maximum likelihood estimate of delay is the solution of

$$\min_\theta [\mathbf{X} - \mathbf{s}(\theta)]^t \mathbf{K}_N^{-1} [\mathbf{X} - \mathbf{s}(\theta)]$$

The term $\mathbf{s}^t\mathbf{K}_N^{-1}\mathbf{s}$ is usually assumed not to vary with the presumed time origin of the signal because of the signal's short duration. If the noise is white, this term is constant except near the "edges" of the observation interval. If not white, the kernel of this quadratic form is equivalent to a whitening filter. As discussed in the previous chapter (§5.4.2 {180}), this filter may be time varying. For noise spectra that are rational and have only poles, the whitening filter's unit-sample response varies only near the edges (see the example {182}). Thus, near the edges, this quadratic form varies with presumed delay and the maximization is analytically difficult. Taking the "easy way out" by ignoring edge effects, the estimate is the solution of

$$\max_{\theta} \left[\mathbf{X}^t\mathbf{K}_N^{-1}\mathbf{s}(\theta)\right]$$

Thus, the delay estimate is the signal time origin that maximizes the matched filter's output.

Cramér-Rao Bound

In addition to the complexity of finding the maximum likelihood estimate, the discrete-valued nature of the parameter also calls into question the use of the Cramér-Rao bound. One of the fundamental assumptions of the bound's derivation is the differentiability of the likelihood function with respect to the parameter. Mathematically, a sequence cannot be differentiated with respect to the integers. A sequence can be differentiated with respect to its argument if we consider the variable to be continuous valued. This approximation can be used only if the sampling interval, unity for the integers, is dense with respect to variations of the sequence. This condition means that the signal must be oversampled to apply the Cramér-Rao bound in a meaningful way. Under these conditions, the mean-squared estimation error for *unbiased estimators* can be no smaller than the Cramér-Rao bound, which is given by

$$\mathsf{E}[\varepsilon^2] \geq \frac{1}{\sum_{k,l}\left[\mathbf{K}_N^{-1}\right]_{k,l}\dot{s}(k-\theta)\dot{s}(l-\theta)}$$

which, in the white-noise case, becomes

$$\mathsf{E}[\varepsilon^2] \geq \frac{\sigma_N^2}{\sum_l [\dot{s}(l)]^2} \tag{4.16}$$

Here, $\dot{s}(\cdot)$ denotes the "derivative" of the discrete-time signal. To justify using this Cramér-Rao bound, we must face the issue of whether an unbiased estimator for time delay *exists*. No general answer exists; each estimator, including the maximum likelihood one, must be examined individually.

Example

Assume that the noise is white. Because of this assumption, we determine the time delay by maximizing the match-filtered observations.

$$\arg\max_{\theta} \sum_l X(l)s(l-\theta) = \widehat{\theta}_{\mathrm{ML}}$$

The number of terms in the sum equals the signal duration. Fig. 4.2 illustrates the match-filtered output in two separate situations; in one the signal has a relatively low-frequency spectrum as compared with the second. Because of the symmetry of the autocorrelation function, the estimate *should* be unbiased so long as the autocorrelation function is completely contained within the observation interval. Direct proof of this claim is left to the masochistic reader. For sinusoidal signals of energy E and frequency ω_0, the Cramér-Rao bound is given by $\mathsf{E}[\varepsilon^2] = \sigma_N^2/\omega_0^2 E$. This bound on the error is accurate only if the measured maximum frequently occurs in the dominant peak of the signal's autocorrelation function. Otherwise, the maximum likelihood estimate "skips" a cycle and produces values concentrated near one of the smaller peaks. The interval between zero crossings of the dominant peak is $\pi/(2\omega_0)$; the signal-to-noise ratio E/σ_N^2 must exceed $4/\pi^2$ (about 0.5). Remember that this result implicitly assumed a low-frequency sinusoid. The second example demonstrates that cycle skipping occurs more frequently than this guideline suggests when a high-frequency sinusoid is used. We have seen that the Cramér-Rao bound does not capture errors related to cycle-skipping well (Figure 4.1 {67}).

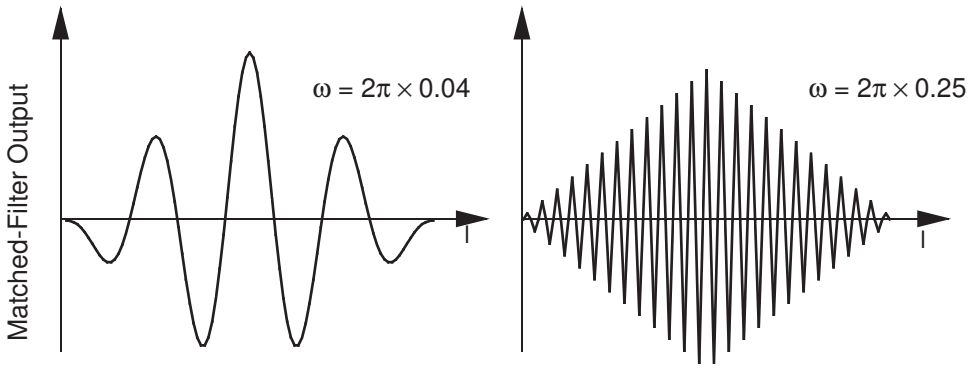


Figure 4.2: The matched filter outputs are shown for two separate signal situations. In each case, the observation interval (100 samples), the signal's duration (50 samples) and energy (unity) are the same. The difference lies in the signal waveform; both are sinusoids with the first having a frequency of $2\pi \times 0.04$ and the second $2\pi \times 0.25$. Each output is the signal's autocorrelation function. Few, broad peaks characterize the low-frequency example whereas many narrow peaks are found in the high frequency one.

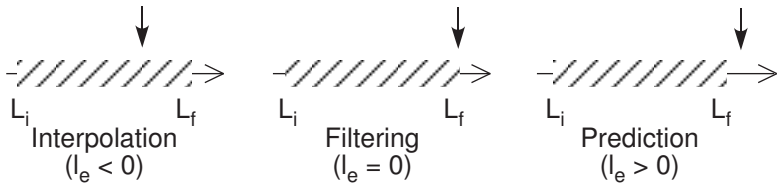


Figure 4.3: The three classical categories of linear signal waveform estimation are defined by the observation interval's relation to the time at which we want to estimate the signal value. As time evolves, so does the observation interval so that l_e, the interval between the last observation and the estimation time, is fixed.

4.4 Linear Signal Waveform Estimation

When the details of a signal's waveform are unknown, describing the signal parametrically is usually unsatisfactory. We need techniques that estimate waveforms rather than numbers. For example, we may want to know the propagating signal's waveform contained in the noise-corrupted array output. Without some *a priori* information, this task is impossible; if neither the signal nor the noise is known, how can anyone discriminate one from the other? The key to waveform estimation is how much prior information we have about the signal and the noise, and how valid that information is. Given noisy observations of a signal throughout the interval $[L_i, L_f]$, the waveform estimation problem is to estimate accurately the value of the signal at some moment $L_f + l_e$. In most situations, the observation interval evolves with the passage of time while the estimation time is fixed relative to the occurrence of the most recent observation (in other words, l_e is a constant). Linear waveform estimation results when we apply a linear filter to the observations.

Waveform estimation problems are usually placed into one of three categories [4: 9–11] based on the value of l_e (see Fig. 4.3):

Interpolation. The interpolation or smoothing problem is to estimate the signal at some moment within the observation interval ($l_e < 0$). Observations are thus considered before and after the time at which the signal needs to be estimated. In practice, applying interpolation filtering means that the estimated signal waveform is produced some time *after* it occurred.

Filtering. We estimate the signal at the end of the observation interval ($l_e = 0$). Thus, a waveform estimate is produced as soon as the signal is observed. The filtering problem arises when we want to remove noise (as much as possible) from noise-corrupted signal observations as they are obtained.

Prediction. Here, we attempt to predict the signal's value at some future time ($l_e > 0$). The signal's structure

must be well known to enable us to predict what values the signal obtains. Prediction filters have obvious applications in sonar/radar tracking and stock market analysis. Of all the waveform estimation problems, this one produces the largest errors.

Waveform estimation algorithms are *not* defined by this categorization; each technique can be applied to each type of problem (in most cases). Instead, the algorithms are defined according to the signal model. Correctness of the signal model governs the utility of a given technique. Because the signal usually appears *linearly* in the expression for the observations (the noise is usually additive), *linear* waveform estimation methods—filters—are frequently employed.

4.4.1 General Considerations

In the context of *linear* waveform estimation, the signal as well as the noise is considered to be a stochastic sequence. Furthermore, the signal component $\widetilde{S}$ in the observations is assumed to only be *related* to the signal s to be estimated and not necessarily equal to it: $X(l) = \widetilde{S}(l) + N(l)$. For example, the observations may contain a filtered version of the signal when we require an estimate of the prefiltered waveform. In this situation, the signal filter is usually known. The noise and signal components are zero-mean random sequences statistically independent of each other. The optimum filter that provides the waveform estimate $\widehat{S}(l)$ can be time invariant (Wiener filters), time varying (Kalman filters), or data dependent (adaptive filters). Choosing an estimation strategy is determined by the signal's characteristics and the degree to which these characteristics are known. For generality, we allow the optimum filter's unit-sample response $h_\diamond(l,k)$ to be time varying: It depends directly on the values of l, the "time variable" and k, the "time" at which the unit sample is presented. When the filter is time invariant, the unit-sample response would be a function of the interval $l-k$, time since presentation of the unit sample. The fundamental form of the observations and the estimated signal in all linear waveform estimators is

$$\boxed{\begin{aligned} X(l) &= \widetilde{S}(l) + N(l) \\ \widehat{S}(L_f + l_e) &= \sum_{k=L_i}^{L_f} h_\diamond(L_f,k) X(k) \end{aligned}}$$

The estimate of the signal's value at $L_f + l_e$ is thus produced at time L_f in the filter's output. The duration of the filter's unit-sample response extends over the entire observation interval $[L_i, L_f]$.

The Orthogonality Principle that proved so useful in linear parameter estimation can be applied here. It states that the estimation error must be orthogonal to all linear transformations of the observations (see Eq. 4.3 {60}). For the waveform estimation problem, this requirement implies that

$$\mathsf{E}\left[\left\{S(L_f+l_e) - \widehat{S}(L_f+l_e)\right\} \sum_{k=L_i}^{L_f} h(L_f,k) X(k)\right] = 0 \quad \text{for all } h(\cdot,\cdot)$$

This expression implies that each observed value must be orthogonal to the estimation error at time $L_f + l_e$.

$$\mathsf{E}\left[\left\{ S(L_f+l_e) - \sum_{j=L_i}^{L_f} h_\diamond(L_f,j) X(j) \right\} X(k)\right] = 0 \quad \text{for all } k \text{ in } [L_i, L_f]$$

Simplifying this expression, the fundamental equation that determines the unit-sample response of the linear minimum mean-squared error filter is

$$\boxed{K_{S\widetilde{S}}(L_f + l_e, k) = \sum_{j=L_i}^{L_f} K_X(j,k) h_\diamond(L_f, j) \quad \text{for all } k \text{ in } [L_i, L_f]}$$

where $K_X(k,l)$ is the covariance function of the observations, equaling $\mathsf{E}[X(k)X(l)]$, and $K_{S\widetilde{S}}(L_f + l_e, k)$ is the cross-covariance between the signal at $L_f + l_e$ and the signal-related component of the observation at k. When

the signal and noise are uncorrelated, $K_X(k,l) = K_{\widetilde{S}}(k,l) + K_N(k,l)$. Given these quantities, the preceding equation must then be solved for the unit-sample response of the optimum filter. This equation is known as the *generalized Wiener-Hopf equation*.

From the general theory of linear estimators, the mean-squared estimation error at index l equals the variance of the quantity being estimated minus the estimate's projection onto the signal.

$$\mathsf{E}[\varepsilon^2(l)] = K_S(l,l) - \mathsf{E}[\widehat{S}(l)S(l)]$$

Expressing the signal estimate as a linear filtering operation on the observations, this expression becomes

$$\mathsf{E}[\varepsilon^2(l)] = K_S(l,l) - \sum_{k=L_i}^{L_f} h_\diamond(L_f,k) K_{S\widetilde{S}}(l,k)$$

Further reduction of this expression is usually problem dependent, as succeeding sections illustrate.

4.4.2 Wiener Filters

Wiener filters are the solutions of the linear minimum mean-squared waveform estimation problem for the special case in which the noise and the signal are *stationary* random sequences [45: 100–18];[89: 481–515];[96]. The covariance functions appearing in the generalized Wiener-Hopf equation thus depend on the difference of their arguments. Considering the form of this equation, one would expect the unit-sample response of the optimum filter to depend on its arguments in a similar fashion. This presumption is in fact valid, and Wiener filters are always time invariant.

$$\widehat{S}(L_f + l_e) = \sum_{k=L_i}^{L_f} h_\diamond(L_f - k) X(k)$$

We consider first the case in which the initial observation time L_i equals $-\infty$. The resulting filter uses all of the observations available at any moment.* The errors that result from using this filter are smaller than those obtained when the filter is constrained to use a finite number of observations (such as some number of recent samples). The choice of $L_i = -\infty$ corresponds to an infinite-duration impulse response (*IIR*) Wiener filter; in a succeeding section, L_i is finite and a finite-duration impulse response (*FIR*) Wiener filter results. The error characteristics of the *IIR* Wiener filter generally bound those of *FIR* Wiener filters because more observations are used. We write the generalized Wiener-Hopf equation for the *IIR* case as

$$K_{S\widetilde{S}}(L_f + l_e - k) = \sum_{j=-\infty}^{L_f} K_X(j-k) \cdot h_\diamond(L_f - j) \quad \text{for all } k \text{ in } (-\infty, L_f]$$

Changing summation variables results in the somewhat simpler expression known as the Wiener-Hopf equation. It and the expression for the mean-squared estimation error are given by

$$\boxed{\begin{aligned} K_{S\widetilde{S}}(l + l_e) &= \sum_{k=0}^{\infty} K_X(l-k) h_\diamond(k) \quad \text{for all } l \text{ in } [0,\infty) \\ \mathsf{E}[\varepsilon^2] &= K_S(0) - \sum_{k=0}^{\infty} h_\diamond(k) K_{S\widetilde{S}}(l_e + k) \end{aligned}} \tag{4.17}$$

The first term in the error expression is the signal variance. The mean-squared error of the signal estimate cannot be greater than this quantity; this error results when the estimate always equals 0.

In many circumstances, we want to estimate the signal directly contained in observations: $X = S + N$. This situation leads to a somewhat simpler form for the Wiener-Hopf equation.

$$K_S(l + l_e) = \sum_{k=0}^{\infty} [K_S(l-k) + K_N(l-k)] h_\diamond(k) \quad \text{for all } l \text{ in } [0,\infty)$$

It is this form we solve, but the previous one is required in its solution.

*Presumably, observations have been continuously available since the beginning of the universe.

Solving the Wiener-Hopf equation. The Wiener-Hopf equation at first glance appears to be a convolution integral, implying that the optimum filter's frequency response could be easily found. The constraining condition—the equation applies only for the variable l in the interval $[0,\infty)$—means, however, that Fourier techniques *cannot* be used for the general case. If the Fourier Transform of the left side of the Wiener-Hopf equation were evaluated only over the constraining interval, the covariance function on the left would be *implicitly* assumed 0 outside the interval, which is usually not the case. Simply stated but mathematically complicated, the covariance function of the signal outside this interval is not to be considered in the solution of the equation.

Non-causal solution. One set of circumstances does allow Fourier techniques. Let the Wiener filter be noncausal with $L_f = +\infty$. In this case, the Wiener-Hopf equation becomes

$$K_S(l) = \sum_{k=-\infty}^{\infty} K_X(l-k)h_\diamond(k) \quad \text{for all } l$$

As this equation must be valid for all values of l, a convolution sum emerges. The frequency response $H_\diamond(f)$ of the optimum filter is thus given by

$$H_\diamond(f) = \frac{\mathscr{S}_S(f)}{\mathscr{S}_S(f)+\mathscr{S}_N(f)}$$

where $\mathscr{S}_S(f)$ and $\mathscr{S}_N(f)$ are, respectively, the signal and the noise power spectra. Because this expression is real and even, the unit-sample response of the optimum filter is also real and even. The filter is therefore noncausal and usually has an infinite duration unit-sample response. This result is not often used in temporal signal processing but may find applications in spatial problems. Be that as it may, because this filter can use the entire set of observations to estimate the signal's value at any moment, it yields the smallest estimation error of *any* linear filter. Computing this error thus establishes a bound on how well any causal or *FIR* Wiener filter performs. The mean-squared estimation error of the noncausal Wiener filter can be expressed in the time domain or frequency domain.

$$\begin{aligned}\mathsf{E}[\varepsilon^2] &= K_S(0) - \sum_{l=-\infty}^{\infty} h_\diamond(l)K_S(l) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\mathscr{S}_S(f)\mathscr{S}_N(f)}{\mathscr{S}_S(f)+\mathscr{S}_N(f)}\,df\end{aligned}$$

Causal solution. To find the causal solution of the Wiener-Hopf equation, we now take L_f to be finite and follow Bode and Shannon's approach [11]. Cleverly note the simplicity of the solution to (4.17) in the seemingly artificial situation where the sum of the signal and noise covariance functions equals a unit sample. In this case, we find that $h_\diamond(l) = K_{S\widetilde{S}}(l+l_e), l = 0,\ldots,\infty$, which is compactly written using step functions as $h_\diamond(l) = K_{S\widetilde{S}}(l+l_e)\mathrm{u}(l)$. In the frequency domain, this result is expressed as

$$H_\diamond(f) = [e^{+j2\pi f l_e}\mathscr{S}_{S\widetilde{S}}(f)]_+,$$

where $[\mathscr{S}(f)]_+$ means the Fourier transform of a covariance function's *additively* causal part, which corresponds to its values at non-negative indices. The "additive" qualifier means we think of the covariance function as the sum of a causal function, its values for non-negative indices, and a non-causal one, the values for negative indices. The artificial situation of unit-sample covariance for the observations can be created by passing them through a *whitening* filter. The optimum filter would then consist of the cascade (conceptually, at least) of the whitening filter and the filter having the unit-sample response equal to the simple solution of the Wiener-Hopf equation just given (Fig. 4.4). The whitening filter is such that the observations become white noise with unity variance. Thus, its frequency response must be equal to the reciprocal of the "square root" of the observation sequence's power spectrum. In the context of this problem, the square root of a given power spectrum $\mathscr{S}(f)$ must result in a frequency response $G(f)$ whose squared magnitude is the solution to

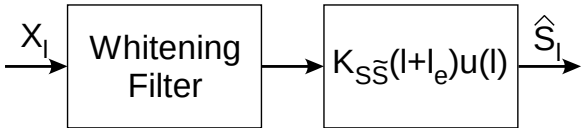


Figure 4.4: We can conceptualize the solution of the Wiener-Hopf equation as the cascade of two linear filters. The first whitens the observations and the second has a unit-sample response equal to (over the appropriate interval) the covariance between the signal S and the signal component $\widetilde{S}$ of the whitened observations.

the problem $G(f)G^*(f) = \mathscr{S}(f)$. Thus, we need the *multiplicatively* causal part of the power density spectrum: in the frequency domain, we decompose the spectrum into a product of two terms, one corresponding to a causal signal and the other to its non-causal relative.

This square-root must be evaluated carefully; for every $\mathscr{S}(f)$, there are an infinity of answers, only some of which are appropriate. For example, the unit-valued squared magnitude has many square roots: not only are the simplistic answers of $+1$ and -1 solutions, but also the frequency responses of allpass filters: filters whose frequency responses are constant in magnitude, but vary in phase. A first-order allpass filter would have a frequency response of the form

$$H(f) = \frac{e^{-j2\pi f} - a}{1 - ae^{-j2\pi f}}.$$

Because of the symmetry of the pole and zero locations (pole at a while the zero is located at $1/a$), this frequency response has a constant magnitude. Factors such as these could be included in the square root of an arbitrary power spectrum, but they are purely ancillary. More importantly, the reciprocal of the power spectrum's square root must be causal. These restrictions allow the number of possible solutions to be reduced to just a few.

Letting the frequency response of the whitening filter be denoted by $H_w(f)$, it must satisfy

$$[\mathscr{S}_S(f) + \mathscr{S}_N(f)] \cdot |H_w(f)|^2 = 1.$$

Consider for the moment rational power spectra that can be expressed as ratios of polynomials in $\exp\{-j2\pi f\}$. Such spectra can envisioned as the result of passing white noise through linear filters having poles and zeros. These power spectra not only have the same poles and zeros as the linear filter, but also at their reciprocals because the power spectrum equals the product of the frequency response *and* its conjugate. For example, when white noise is passed through a linear filter having transfer function $G(z) = (1 - bz^{-1})/(1 - az^{-1})$, the power spectrum of the output sequence equals $G(f)G^*(f)$, which becomes

$$\begin{aligned}\mathscr{S}(f) &= \frac{\left(1 - be^{-j2\pi f}\right)\left(1 - be^{+j2\pi f}\right)}{\left(1 - ae^{-j2\pi f}\right)\left(1 - ae^{+j2\pi f}\right)} \\ &= \frac{b}{a}\frac{\left(1 - be^{-j2\pi f}\right)\left(1 - \frac{1}{b}e^{-j2\pi f}\right)}{\left(1 - ae^{-j2\pi f}\right)\left(1 - \frac{1}{a}e^{-j2\pi f}\right)}.\end{aligned}$$

Given a power spectrum such as this and wanting to derive a frequency response corresponding to the stable, causal filter which could have generated it, the poles located inside the unit circle and *either* of the zeros occurring in reciprocal pairs could be chosen. Usually the selected zeros are those located inside the unit circle; in this way, the so-called *minimum-phase* square root is chosen. We have thus calculated the square-root of a power spectrum! The result of finding the stable, causal, and minimum-phase square-root of $\mathscr{S}(f)$ is denoted by $\mathscr{S}^+(f)$. The whitening filter's frequency response equals the reciprocal of this quantity.

$$H_w(f) = \frac{1}{[\mathscr{S}_S + \mathscr{S}_N]^+(f)}$$

Note that the zeros located inside the unit circle must have been selected in the square root extraction lest the whitening filter be unstable!

Once the observations are whitened, the solution to the simplified Wiener-Hopf equation given earlier is related to the *whitened* signal component. The optimum filter's unit-sample response corresponds to the values of the cross-covariance between the signal and its whitened version evaluated over $[l_e, \infty)$.

$$\begin{aligned} K_{S\widetilde{S}}(l) &= \mathsf{E}[S(k)\widetilde{S}(k-l)] \\ &= \mathsf{E}\left[S(k)\sum_{j=-\infty}^{\infty} h_w(j)S(k-l-j)\right] \\ &= \sum_{j=-\infty}^{\infty} h_w(-j)K_S(l-j) \\ K_{S\widetilde{S}}(l) &= h_w(-l)\otimes K_S(l) \\ \Longrightarrow \mathscr{S}_{S\widetilde{S}}(f) &= \mathscr{S}_S(f)H_w^*(f) \end{aligned}$$

In the frequency domain, the selection of the index values is thus equivalent to extracting the causal part of $\exp\{+j2\pi f l_e\}\mathscr{S}_S(f)H_w^*(f)$. However, the causal part is evaluated in a different sense from that used in constructing the whitening filter. Here, we want that aspect of the power spectrum which corresponds to the causal part in the *additive* sense while in deriving the whitening filter we sought the causal part in a *convolutional* sense (i.e., factorization of the spectrum into causal and non-causal components). To compute the additively causal part of a power spectrum, the power spectrum must be expanded in a partial fraction expansion with those terms corresponding to poles inside the unit circle selected. This procedure results in the operation previously denoted by $[\mathscr{S}(f)]_+$.

The solution to the Wiener-Hopf equation—the frequency response of the Wiener filter—is the product of the frequency response of the whitening filter and the frequency response of the signal estimation filter based on whitened observations.

$$\boxed{H_\diamond(f) = \frac{1}{[\mathscr{S}_S+\mathscr{S}_N]^+(f)}\left[\frac{e^{+j2\pi f l_e}\mathscr{S}_S(f)}{[\mathscr{S}_S+\mathscr{S}_N]^{+*}(f)}\right]_+}$$

Example

Let's estimate the value of $S(L_f+l_e)$ with a Wiener filter using the observations obtained up to and including time L_f. The additive noise in the observations is white, having variance $8/7$. The power spectrum of the signal is given by

$$\begin{aligned} \mathscr{S}_S(f) &= \frac{1}{5/4-\cos 2\pi f} \\ &= \frac{1}{1-0.5e^{-j2\pi f}}\cdot\frac{1}{1-0.5e^{+j2\pi f}} \end{aligned}$$

The variance of the signal equals the value of the covariance function (found by the inverse Fourier Transform of this expression) at the origin. In this case, the variance equals $4/3$; the signal-to-noise ratio of the observations, taken to be the ratio of their variances, equals $7/6$.

The power spectrum of the observations is the sum of the signal and noise power spectra.

$$\begin{aligned} \mathscr{S}_S(f)+\mathscr{S}_N(f) &= \frac{1}{1-0.5e^{-j2\pi f}}\frac{1}{1-0.5e^{+j2\pi f}}+\frac{8}{7} \\ &= \frac{16}{7}\frac{\left(1-0.25e^{-j2\pi f}\right)\left(1-0.25e^{+j2\pi f}\right)}{\left(1-0.5e^{-j2\pi f}\right)\left(1-0.5e^{+j2\pi f}\right)} \end{aligned}$$

The noncausal Wiener filter has the frequency response

$$H_\diamond(f) = \frac{\mathscr{S}_S(f)}{\mathscr{S}_S(f)+\mathscr{S}_N(f)} = \frac{7}{16}\frac{1}{\left(1-0.25e^{-j2\pi f}\right)\left(1-0.25e^{+j2\pi f}\right)}$$

The unit-sample response corresponding to this frequency response and the covariance function of the signal are found to be

$$h_\diamond(l) = \frac{7}{15}\left(\frac{1}{4}\right)^{|l|} \quad \text{and} \quad K_S(l) = \frac{4}{3}\left(\frac{1}{2}\right)^{|l|}$$

We find that the mean-squared estimation error for the noncausal estimator equals $4/3 - 4/5 = 8/15$.

The convolutionally causal part of signal-plus-noise power spectrum consists of the first terms in the numerator and denominator of the signal-plus-noise power spectrum.

$$[\mathscr{S}_S + \mathscr{S}_N]^+(f) = \frac{4}{\sqrt{7}}\frac{1-0.25e^{-j2\pi f}}{1-0.5e^{-j2\pi f}}$$

The second term in the expression for the frequency response of the optimum filter is given by

$$\frac{e^{+j2\pi f l_e}\mathscr{S}_S(f)}{[\mathscr{S}_S + \mathscr{S}_N]^{+*}(f)} = \frac{\dfrac{e^{+j2\pi f l_e}}{\left(1-0.5e^{-j2\pi f}\right)\left(1-0.5e^{+j2\pi f}\right)}}{\dfrac{4}{\sqrt{7}}\dfrac{1-0.25e^{+j2\pi f}}{1-0.5e^{+j2\pi f}}}$$

$$= \frac{\sqrt{7}}{4}\frac{e^{+j2\pi f l_e}}{(1-0.5e^{-j2\pi f})(1-0.25e^{+j2\pi f})}$$

The additively causal part of this Fourier Transform is found by evaluating its partial fraction expansion.

$$\frac{\sqrt{7}}{4}\frac{e^{+j2\pi f l_e}}{(1-0.5e^{-j2\pi f})(1-0.25e^{+j2\pi f})} = \frac{e^{+j2\pi f l_e}}{2\sqrt{7}}\left[\frac{4}{1-0.5e^{-j2\pi f}} - \frac{2e^{+j2\pi f}}{1-0.25e^{+j2\pi f}}\right]$$

The simplest solution occurs when l_e equals zero: Estimate the signal value at the moment of the most recent observation. The first term on the right side of the preceding expression corresponds to the additively causal portion.

$$\left[\frac{\mathscr{S}_S(f)}{[\mathscr{S}_S + \mathscr{S}_N]^{+*}(f)}\right]_+ = \frac{2}{\sqrt{7}}\frac{1}{1-0.5e^{-j2\pi f}}$$

The frequency response of the Wiener filter is found to be

$$H_\diamond(f) = \frac{\sqrt{7}}{4}\frac{1-0.5e^{-j2\pi f}}{1-0.25e^{-j2\pi f}}\cdot\frac{2}{\sqrt{7}}\frac{1}{1-0.5e^{-j2\pi f}}$$

$$= \frac{1}{2}\frac{1}{1-0.25e^{-j2\pi f}}$$

The Wiener filter has the form of a simple first-order filter with the pole arising from the whitening filter. The difference equation corresponding to this frequency response is

$$\widehat{S}(l) = \frac{1}{4}\widehat{S}(l-1) + \frac{1}{2}X(l)$$

The waveforms that result in this example are exemplified in Fig. 4.5.

To find the mean-squared estimation error, we subtract from the signal's variance the summed product of the Wiener filter's unit-sample response and the cross-covariance between S and $\widetilde{S}$ (4.17). The cross-covariance equals the inverse transform of the Wiener filter's second component and therefore equals $(2/\sqrt{7})(1/2)^l$, $l \geq 0$. The mean-squared estimation error is numerically equal to

$$\mathsf{E}[\varepsilon^2] = \frac{4}{3} - \sum_{l=0}^{\infty}\frac{1}{2}\left(\frac{1}{4}\right)^l \cdot \frac{2}{\sqrt{7}}\left(\frac{1}{2}\right)^l$$

$$= \frac{4}{3} - \frac{1}{\sqrt{7}}\sum_{l=0}^{\infty}\left(\frac{1}{8}\right)^l$$

$$= \frac{4}{3} - \frac{8}{7\sqrt{7}} = 0.9$$

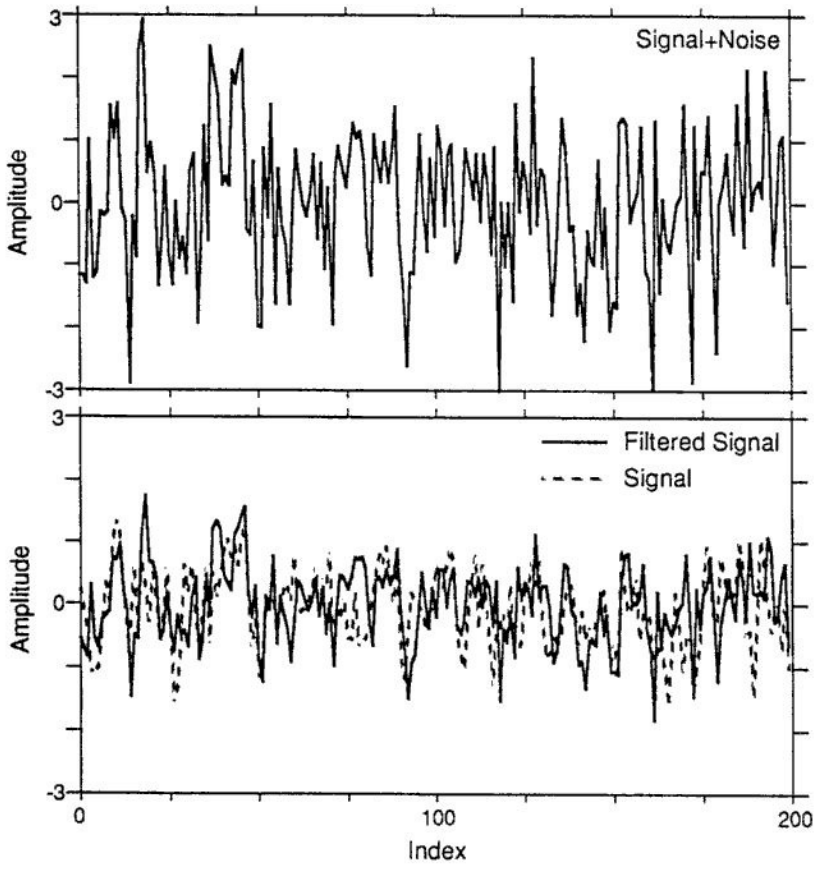


Figure 4.5: The upper panel displays observations having statistic characteristics corresponding to those given in the accompanying example. The output of the causal Wiener filter is shown in the bottom panel along with the actual signal, which is shown as a dashed line.

Compared with the smallest possible value of 0.53 provided by the noncausal Wiener filter, using a causal filter yields about a factor of two larger mean-squared error. The signal-to-noise ratio of the estimated signal is equal to $K_S(0)/\mathsf{E}[\varepsilon^2]$. The causal filter yields a signal-to-noise ratio of 1.48, which should be compared with the ratio of 1.17 in the observations. The ratio of the signal-to-noise ratios at the output and input of a signal processing operation is usually referred to as the *processing gain*. The best possible processing gain is 2.14 and equals 1.11 in the causal case. These rather modest gains are due to the close similarity between the power spectra of the signal and the noise. As the parameter of the signal's power spectrum is increased, the two become less similar, and the processing gain increases.

Now consider the case in which $l_e > 0$: We want to predict the signal's future value. The whitening filter portion of the solution does not depend on the value of l_e and is therefore identical to that just given. The second component of the Wiener filter's frequency response does depend on l_e and is given for positive values of l_e by

$$\left[\frac{e^{+j2\pi f l_e}\mathscr{S}_S(f)}{[\mathscr{S}_S+\mathscr{S}_N]^{+*}(f)}\right]_+ = \left[\frac{\frac{2}{\sqrt{7}}e^{+j2\pi f l_e}}{1-0.5e^{-j2\pi f}}\right]_+$$

The causal portion of this frequency response is found by shifting the unit-sample response to the *left* and retaining the positive-time portion. Because this frequency response has only one pole, this

manipulation is expressed simply as a scaling.

$$\left[\frac{e^{+j2\pi f l_e}\mathscr{S}_S(f)}{[\mathscr{S}_S+\mathscr{S}_N]^{+*}(f)}\right]_+ = \frac{2}{\sqrt{7}}\frac{\left(\frac{1}{2}\right)^{l_e}}{1-0.5e^{-j2\pi f}}$$

The frequency response of the prediction filter is thus given by

$$H_\diamond(f) = \frac{1}{2}\frac{2^{-l_e}}{1-0.25e^{-j2\pi f}}$$

The optimum linear predictor is a *scaled* version of the signal estimator. The mean-squared error increases as the desired time of the predicted value exceeds the time of the last observation. In particular, the signal-to-noise ratio of the predicted value is given by

$$\frac{K_S(0)}{\mathsf{E}[\varepsilon^2]} = \frac{1}{1-\frac{4}{7}\left(\frac{1}{2}\right)^{2l_e}}$$

The signal-to-noise ratio decreases rapidly as the prediction time extends into the future. This decrease is directly related to the reduced correlation between the signal and its future values in this example. This correlation is described by the absolute value of the signal's covariance function relative to its maximum value at the origin. As a covariance function broadens (corresponding to a lower frequency signal), the prediction error decreases. If a covariance function oscillates, the mean-squared prediction error varies in a similar fashion.

Finite-duration Wiener filters. Another useful formulation of Wiener filter theory is to constrain the filter's unit-sample response to have finite duration. To find this solution to the Wiener-Hopf equation, the values of L_f and L_i are chosen to be finite. Letting L represent the duration of the filter's unit-sample response ($L = L_f - L_i + 1$), the Wiener-Hopf equation becomes

$$K_{S\widetilde{S}}(l+l_e) = \sum_{k=0}^{L-1} K_X(k-l)h_\diamond(k), \quad \text{for all } l \text{ in } [0, L-1]$$

This system of equations can be written in matrix form as $\mathbf{k}_{S\widetilde{S}}(l_e) = \mathbf{K}_X\mathbf{h}_\diamond$.

$$\begin{bmatrix} K_{S\widetilde{S}}(l_e) \\ K_{S\widetilde{S}}(l_e+1) \\ \vdots \\ K_{S\widetilde{S}}(l_e+L-1) \end{bmatrix} = \begin{bmatrix} K_X(0) & K_X(1) & \cdots & K_X(L-1) \\ K_X(-1) & K_X(0) & \cdots & K_X(L-2) \\ \vdots & K_X(-1) & \ddots & \vdots \\ K_X(-L+1) & \cdots & K_X(-1) & K_X(0) \end{bmatrix} \cdot \begin{bmatrix} h_\diamond(0) \\ h_\diamond(1) \\ \vdots \\ h_\diamond(L-1) \end{bmatrix}$$

When the signal component of the observations equals the signal being estimated ($\widetilde{S} = S$), the Wiener-Hopf equation becomes $\mathbf{k}_S(l_e) = \mathbf{K}_X\mathbf{h}_\diamond$. The $L \times L$ matrix $\mathbf{K}_X$ is the covariance matrix of the sequence of L observations. In the simple case of uncorrelated signal and noise components, this covariance matrix is the sum of those of the signal and the noise ($\mathbf{K}_X = \mathbf{K}_S + \mathbf{K}_N$). This matrix has an inverse in all but unusual circumstances with the result that the unit-sample response of the *FIR* Wiener filter is given by

$$\boxed{\mathbf{h}_\diamond = \mathbf{K}_X^{-1}\mathbf{k}_S(l_e)}$$

Because this covariance matrix is Toeplitz and Hermitian, its inverse can be efficiently computed using a variety of algorithms [71: 80–90]. The mean-squared error of the estimate is given by

$$\begin{aligned}\mathsf{E}[\varepsilon^2] &= K_S(0) - \sum_{k=0}^{L-1} h_\diamond(k)K_S(l_e+k) \\ &= K_S(0) - \mathbf{k}_S^t(l_e)\mathbf{K}_X^{-1}\mathbf{k}_S(l_e)\end{aligned}$$

Linear prediction. One especially important variation of the *FIR* Wiener filter occurs in the unique situation in which no observation noise is present, and the signal generation model contains only poles [69, 70]. Thus, the signal $S(l)$ is generated by passing white noise $w(l)$ through a linear system given by the difference equation

$$S(l) = a_1 S(l-1) + a_2 S(l-2) + \cdots + a_p S(l-p) + w(l)$$

The coefficients $a_1, \ldots, a_p$ are unknown. This signal modeling approach is frequently used to estimate the signal's spectrum.

As no noise is present in the observations, the filtered estimate of the signal ($l_e = 0$) equals $S(l)$ and the estimation error is exactly 0. The concern of linear prediction is not this trivial problem, but the so-called *one-step prediction problem* ($l_e = 1$): Predict the value of the signal at index l given values of $S(l-1)$,$S(l-2), \ldots$. Thus, we seek a *FIR* Wiener filter predictor, which has the form

$$\widehat{S}(l) = h(0)S(l-1) + h(1)S(l-2) + \cdots + h(p-1)S(l-p)$$

Comparing the signal model equation to that for the Wiener filter predictor, we see that the model parameters $\{a_1, \ldots, a_p\}$ equal the Wiener filter's unit-sample response $h(\cdot)$ because the input $w(l)$ is uncorrelated from sample to sample. In linear prediction, the signal model parameters are used notationally to express the filter coefficients.

The Orthogonality Principle can be used to find the minimum mean-squared error predictor of the next signal value. By requiring orthogonality of the prediction error to each of the observations used in the estimate, the following set of equations results.

$$\begin{aligned}
a_1 K_S(0) + a_2 K_S(1) + \cdots + a_p K_S(p-1) &= K_S(1) \\
a_1 K_S(1) + a_2 K_S(0) + \cdots + a_p K_S(p-2) &= K_S(2) \\
\vdots \qquad\qquad\qquad\qquad\qquad\qquad & \qquad \vdots \\
a_1 K_S(p-1) + a_2 K_S(p-2) + \cdots + a_p K_S(0) &= K_S(p)
\end{aligned}$$

In linear prediction, these are known as the *Yule-Walker equations*. Expressing them concisely in matrix form $\mathbf{K}_S\mathbf{a} = \mathbf{k}_S$, the solution is $\mathbf{a} = \mathbf{K}_S^{-1}\mathbf{k}_S$.

From the signal model equation, we see that the mean-squared prediction error $\mathsf{E}[\{S(l) - \widehat{S}(l)\}^2]$ equals the variance σ_w^2 of the white-noise input to the model. Computing the mean-squared estimation error according to Eq. 4.17 {85}, this variance is expressed by

$$\sigma_w^2 = K_S(0) - a_1 K_S(1) - \cdots - a_p K_S(p)$$

This result can be combined with the previous set of equations to yield a unified set of equations for the unknown parameters and the mean-squared error of the optimal linear predictive filter.

$$\begin{bmatrix} K_S(0) & K_S(1) & \cdots & K_S(p) \\ K_S(1) & K_S(0) & \cdots & K_S(p-1) \\ \vdots & K_S(1) & \ddots & \vdots \\ K_S(p) & \cdots & K_S(1) & K_S(0) \end{bmatrix} \cdot \begin{bmatrix} 1 \\ -a_1 \\ \vdots \\ -a_p \end{bmatrix} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.18}$$

To solve this set of equations for the model coefficients and the input-noise variance conceptually, we compute the preliminary result $\mathbf{K}_S^{-1}\boldsymbol{\delta}$. The first element of this vector equals the reciprocal of σ_w^2; normalizing $\mathbf{K}_S^{-1}\boldsymbol{\delta}$ so that its leading term is unity yields the coefficient vector $\mathbf{a}$. Levinson's algorithm can be used to solve these equations efficiently and simultaneously obtain the noise variance [71: 211–16].

4.4.3 Dynamic Adaptive Filtering

The Wiener filter approach presumes that the signal *and* the noise are statistically well characterized. Specifically, the theoretical framework requires a second-order description of the signal and the noise that models

the signal and the noise as white noise processes that have been passed through *known* linear system. Each filter has a specific implementation and performance characteristics that must be determined before the observations are made.

In many other situations, however, the signal or the observation noise are ill defined. They may vary in unpredictable ways or have poorly specified spectral or temporal characteristics. One typical problem is *interference*: Rather than just noise, the observations are corrupted by other signals that defy simple statistical characterization. Here, the waveform estimator must *learn* from the observations what is not known *a priori* and *adapt* its characteristics or structure as data become available. Rather than, for example, estimating a signal's power spectrum "off-line" and then incorporating the estimate into the filter's structure for subsequent observations, *dynamic adaptive filters* either take the imprecise knowledge into account directly or modify themselves while producing signal estimates [3, 24, 45, 47, 95].

Such filters can adjust two aspects of themselves: their structure and their coefficients. Consider a finite-duration (*FIR*) Wiener filter. The characteristics of this filter are its duration (L) and the values of the unit-sample response $\mathbf{h}$. The quantity L is a structural parameter because it determines the number of filter coefficients; the coefficients are determined by signal and noise characteristics. Structure adjustment is usually quite involved while changing coefficient values dynamically is much more plausible. For example, adjusting the filter coefficients from one sample to another by a presumably modest increment poses no fundamental difficulties (as shown in sequel). Modifying the number of coefficients, however, requires the ability to compute an initial value for a new coefficient, conceptually equivalent to "adjusting" the value from zero. Not only is this adjustment in reality a rather drastic change, but it also increases the computations required for the other coefficients. Consequently, we concentrate here on coefficient adaptations to changes in the observations.

The *FIR* Wiener filter consists of a finite set of coefficients that minimize the mean-squared estimation error. The closed form solution for this set is

$$\mathbf{h}_{\diamond} = \mathbf{K}_X^{-1}\mathbf{k}_{S\widetilde{S}}(l_e)$$

where $\mathbf{K}_X$ is the covariance matrix of the observations (assumed to consist of signal plus noise) and $\mathbf{k}_{S\widetilde{S}}(l_e)$ consists of L values starting at lag l_e of the cross-covariance function between the signal and its observed counterpart. For simplicity, only the one-step prediction problem ($l_e = 1$) is explicitly considered here. To determine a method of *adapting* the filter coefficients at each sample, two approaches have been used. The first one considered here, the least mean squares (*LMS*) algorithm, is to return to the linear minimum mean-squared error minimization problem and find iterative methods of solving it that can incorporate new observations "on the fly." The second approach, the recursive least squares (*RLS*) algorithm (considered in the next section), amounts to finding ways of updating the optimal Wiener solution at each index.

Least mean squares (LMS) algorithm. The *LMS* algorithm is intimately related to an optimization method known as the *method of steepest descent* [3: 46–52];[27]: Given a function to be minimized, the mean-squared error $\mathsf{E}[\varepsilon^2]$ at index l, and the independent variables, the vector of filter coefficients $\mathbf{h}$, the minimum can be found iteratively by adjusting the filter coefficients according to the gradient of the mean-squared error.

$$\mathbf{h}(i+1) = \mathbf{h}(i) - \mu\,\nabla_h\,\mathsf{E}[\varepsilon^2(l)]\,, \quad \mu > 0$$

The gradient "points" in the direction of the maximum rate of change of the mean-squared error for any set of coefficient values. The direction of the gradient's negative points toward the minimum. The often-used analogy is rolling a ball down the edge of a bowl. The direction the ball rolls at each instant corresponds to the locally maximum slope, the negative of the gradient at each spatial location. If the bowl is smooth and isn't too bumpy, the ball rolls to the global minimum; if bumpy, the ball may stop in a local minimum. Because of the characteristics of the mean-squared error, the bowl it forms is provably convex; hence, it has only one minimum. The method of steepest descent is therefore guaranteed to converge.

To compute this gradient, the mean-squared error must be expressed in terms of the filter coefficients. Letting the previous L observations be expressed by the observation vector $\mathbf{X} = \mathrm{col}[X(l-1), \ldots, X(l-L)]$

with $\mathbf{X} = \widetilde{\mathbf{S}} + \mathbf{N}$, the application of an arbitrary set of coefficients results in a mean-squared error given by

$$\begin{aligned}\mathsf{E}[\varepsilon^2] &= \mathsf{E}[(S - \mathbf{h}^t\mathbf{X})^2] \\ &= \sigma_S^2 - 2\mathbf{h}^t\mathbf{k}_{S\widetilde{S}} + \mathbf{h}^t\mathbf{K}_X\mathbf{h}\end{aligned}$$

The gradient of this quantity is, therefore,

$$\nabla_h \mathsf{E}[\varepsilon^2] = -2\mathbf{k}_{S\widetilde{S}} + 2\mathbf{K}_X\mathbf{h}$$

and the method of steepest descent is characterized by the iterative equation for the filter coefficients given by*

$$\mathbf{h}(i+1) = \mathbf{h}(i) + 2\mu[\mathbf{k}_{S\widetilde{S}} - \mathbf{K}_X\mathbf{h}(i)]$$

Here, the variable i denotes iteration number, which in general need not be related to index values: Filter adaptation could occur every once in a while rather than at every index. The term in the brackets can also be written as

$$\mathbf{k}_{S\widetilde{S}} - \mathbf{K}_X\mathbf{h}(i) = \mathsf{E}[\varepsilon(i)\mathbf{X}]$$

where $\varepsilon(i) = S - \mathbf{h}^t(i)\mathbf{X}$ is the estimation error.

The convergence of the method of steepest descent is controlled by the value of the *adaptation parameter* μ. The difference of the filter coefficients from the optimal set, $\mathbf{d}(i) = \mathbf{h}_\diamond - \mathbf{h}(i)$, is governed by an easily derived iterative equation.

$$\begin{aligned}\mathbf{h}_\diamond - \mathbf{h}(i+1) &= \mathbf{h}_\diamond - \big\{\mathbf{h}(i) + 2\mu[\mathbf{k}_{S\widetilde{S}} - \mathbf{K}_X\mathbf{h}(i)]\big\} \\ \mathbf{d}(i+1) &= \mathbf{d}(i) - 2\mu[\mathbf{k}_{S\widetilde{S}} - \mathbf{K}_X\mathbf{h}(i)] \\ \mathbf{d}(i+1) &= \mathbf{d}(i) - 2\mu\mathbf{K}_X[\mathbf{K}_X^{-1}\mathbf{k}_{S\widetilde{S}} - \mathbf{h}(i)] \\ \mathbf{d}(i+1) &= \mathbf{d}(i) - 2\mu\mathbf{K}_X[\underbrace{\mathbf{h}_\diamond - \mathbf{h}(i)}_{\mathbf{d}(i)}] \\ \mathbf{d}(i+1) &= [\mathbf{I} - 2\mu\mathbf{K}_X]\mathbf{d}(i)\end{aligned}$$

This equation says that the error at each iteration is "proportional" to the error at the previous iteration; consequently, the method of steepest descent converges linearly. A closed form expression for the vector of coefficient differences is

$$\mathbf{d}(i) = [\mathbf{I} - 2\mu\mathbf{K}_X]^i\mathbf{d}(0)$$

Convergence—$\lim_{i\to\infty}\mathbf{d}(i) = \mathbf{0}$—is assured *for any initial condition* $\mathbf{d}(0)$ so long as the i^{th} power of the matrix in brackets decreases to 0 as i becomes large. This condition is more easily expressed in terms of the eigenvalues of this matrix (see §B.5 {242}): A necessary and sufficient condition for convergence is that the magnitude of the eigenvalues of $\mathbf{I} - 2\mu\mathbf{K}_X$ must all be less than unity (see Problem 4.36 {142}). Coupling this condition with the necessity that $\mu > 0$ results in the requirement for the method of steepest descent to converge, μ must be less than the reciprocal of the largest eigenvalue of $\mathbf{K}_X$.

$$\boxed{0 < \mu < \frac{1}{\max\lambda_{\mathbf{K}_X}}}$$

The "time constant" of the convergence is governed by the *smallest* eigenvalue and is greater than $1/(2\mu\min\{\lambda_{\mathbf{K}_X}\})$. The greatest rate of convergence is obtained with the smallest time constant; we should thus choose μ as near to its upper limit as possible. When the largest value of μ is selected, the rate of convergence of the method of steepest descent is proportional to the ratio of the largest and smallest eigenvalues of the observations' covariance matrix.

The preceding discussion described a numerical procedure for computing the coefficients of the *FIR* Wiener filter as an alternative to the closed-form expression that involves a matrix inverse. This procedure

*The form of this equation is remarkably similar to a Kalman filter equation for estimating the filter coefficients. This similarity is no accident and reveals the fundamental importance of the ideas expressed by Kalman filter.

does *not* yield an adaptive filter because we must somehow calculate $\mathsf{E}[\varepsilon(i)\mathbf{X}]$, but it does inspire an idea. Expressing the steepest descent iteration in terms of the estimation error yields

$$\mathbf{h}(i+1) = \mathbf{h}(i) + 2\mu\,\mathsf{E}[\varepsilon(i)\mathbf{X}]$$

The *LMS* algorithm amounts to approximating the expected value by the product of the observation vector and the estimation error, a technique known as *stochastic approximation* [91]. As unworkable as it might appear, the expected value is replaced by its *one-sample* average value.

$$\mathsf{E}[\varepsilon(i)\mathbf{X}] \longrightarrow \varepsilon(i)\mathbf{X}$$

The optimization iteration now coincides with the index of the observations and a desired signal ($i = l$). The equations governing the *LMS* algorithm are [3: 68–83];[45: 194–259]

$$\boxed{\begin{aligned} \widehat{S}(l) &= \sum_{k=0}^{L-1} h(l,k)X(l-k-1) = \mathbf{h}^t(l)\mathbf{X}(l-1) \\ \varepsilon(l) &= S(l) - \widehat{S}(l) \\ \mathbf{h}(l+1) &= \mathbf{h}(l) + 2\mu\varepsilon(l)\mathbf{X}(l-1) \end{aligned}}$$

Employing the *LMS* algorithm presumes that either the signal is known to some degree and we want to estimate some aspect of it or that we want to cancel it so that the error sequence $\varepsilon(l)$ is "signal-free" [3: 87–98];[45: 7–31]. A typical example is one-step prediction where no observation noise is present: The filter predicts the signal value at index l, the error at time l is computed, and this error is used to adapt the filter coefficients that yield the *next* signal estimate. Because of the limited extent of the data used to change the filter coefficients, we can estimate well signals having slowly varying characteristics.

Because the filter coefficients depend on new observations and are continually updated via the estimate of the gradient, the *LMS* filter coefficients do *not* converge to a stable set of values even when the observations are stationary. Rather, after an initial transient response to the onset of the observations, the coefficients "settle" toward a set of values and vary randomly to some degree. Furthermore, the convergence bounds on the adaptation parameter must be sharpened to include the gradient estimate rather than its expected value [36]. For the *LMS* filter to converge, the adaptation parameter must satisfy [35, 36]

$$0 < \sum_j \frac{\mu\lambda_j}{1-2\mu\lambda_j} < 1 \quad \text{and} \quad 0 < \mu < \frac{1}{2\max_j\{\lambda_j\}}$$

where $\lambda_j, j = 1,\ldots,L$ denotes the eigenvalues of $\mathbf{R}_X$. Enforcing this condition requires clairvoyance because the covariance is assumed to be unknown. A more practical and more stringent bound is [35]

$$0 < \mu \leq \frac{1}{3\,\mathrm{tr}[\mathbf{R}_X]} \tag{4.19}$$

because the trace of the covariance matrix equals the total power in the observations, an easily estimated quantity in practice.

The variances of the filter's coefficients are also related to the adaptation parameter μ, but in the opposite way than the settling time. A small value of μ results in little change of the coefficients from one update to the next; the filter coefficients settle slowly but their long-term values do not vary greatly about the optimal values. Larger values of μ allow greater changes; more rapid settling occurs, but the long-term values have larger variations. Fig. 4.6 illustrates this point. The degree of the long-term variations in stationary environments is assessed by the so-called *misadjustment factor M*: the difference between the variances of the *LMS* estimate and the optimal, Wiener filter estimate normalized by the optimal value [45: 236–37].

$$M = \frac{\sigma_\varepsilon^2 - \sigma_\varepsilon^2(Wiener)}{\sigma_\varepsilon^2(Wiener)}$$

The misadjustment factor M_{LMS} for the *LMS* algorithm is well approximated (for small values of the factor) by $\mu L\sigma_S^2$ [36];[45: 236–37].

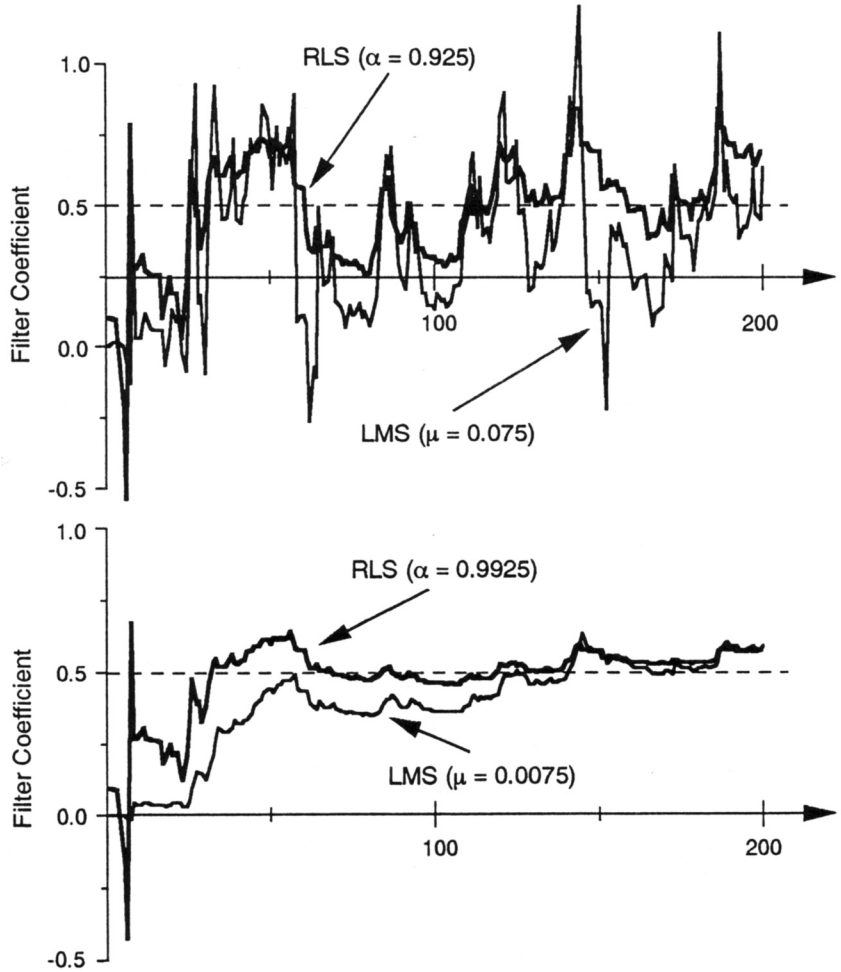


Figure 4.6: The filter coefficients of one-step predictors produced by first-order *LMS* and *RLS* algorithms are shown as a function of the number of observations. The coefficient's correct value is $1/2$. In the upper panel, both algorithms settle rapidly but vary greatly about their optimal values. In the lower, the algorithms' parameters were adjusted to reduce the variations of their estimates. Note that the settling time of the *RLS* filter coefficient is not strongly influenced by this parameter value change. In comparison, settling takes much longer in the *LMS* algorithm.

On the basis of these simple considerations, one is tempted to consider changing the adaptation parameter in addition to the filter coefficients; conceptually, large values could be used initially and then slowly reduced to yield smaller coefficient variations about their optimal values. Given sufficient detail about the signal characteristics, such a scheme might be feasible. If this reduction of μ is tried in an *ad hoc* manner, however, the coefficients might settle about artificial values because they are not allowed to change.*

Recursive least squares (RLS) algorithm. Another approach in adaptive filtering is to approximate the solution for optimal set of filter coefficients *directly* without imposing the additional burden of approximating an optimization procedure as the *LMS* algorithm does [3: 111–21];[36];[45: 381–444]. This more direct approach results in filters having better convergence properties but at the expense of computational complexity.

*Note that *any* set of vectors is a solution to $\mathbf{h}(l+1) = \mathbf{h}(l)$ when $\mu = 0$.

Recalling the solution to the *FIR* Wiener filter, the filter coefficients are given by

$$\mathbf{h}_\diamond = \mathbf{K}_X^{-1}\mathbf{k}_{S\widetilde{S}}$$

The data-dependent quantity in this expression is the covariance matrix of the observations. Not only is a recursive method of computing the inverse of this matrix sought, but also a recursive estimate of the matrix itself. To estimate the covariance matrix of size $L\times L$ from zero-mean data, the data is formed into a series of groups (frames) of duration L. The l^{th} frame, denoted by the vector $\mathbf{X}(l)$, contains the most recent L observations: $\mathbf{X}(l) = \text{col}[X(l), X(l-1), \ldots, X(l-L+1)]$. Two methods of estimating the covariance matrix are commonly used. Whichever is chosen, the estimate at index l must depend on*only* the observations obtained *before* that time.* The first method of estimating the covariance matrix is to average the outer products computed from the most recent S frames.

$$\widehat{\mathbf{K}}_X(l) = \frac{1}{S}\sum_{k=l-S}^{l-1}\mathbf{X}(k)\mathbf{X}^t(k)$$

Note that we must require $S \geq L$ to obtain an invertible covariance matrix from this *moving average* or *Bartlett* estimate. Updating this estimate requires storing all of the $S+L-1$ observations. A second and somewhat simpler approach is to use the *exponentially weighted* estimate†

$$\widehat{\mathbf{K}}_X(l) = \sum_{k=0}^{l-1}\alpha^{l-1-k}\mathbf{X}(k)\mathbf{X}^t(k)$$

This estimate is reminiscent of the convolution sum for a first-order recursive filter having the sequence of outer products for each frame as its input. Because of this analogy, this estimate's promised simplicity arises because it can be written in the recursive form

$$\widehat{\mathbf{K}}_X(l) = \alpha\widehat{\mathbf{K}}_X(l-1) + \mathbf{X}(l-1)\mathbf{X}^t(l-1)$$

Similar to the moving average estimate, an invertible estimate emerges once $l \geq S$. Because of this recursive form, the exponentially weighted estimate has the additional advantage that its inverse can also be expressed recursively. Based on the well-known Matrix Inverse Lemma (see §B.3 {239}), the inverse of $\widehat{\mathbf{K}}_X(l)$ is thus computed according to

$$\widehat{\mathbf{K}}_X^{-1}(l) = \frac{1}{\alpha}\left[\widehat{\mathbf{K}}_X^{-1}(l-1) - \frac{\widehat{\mathbf{K}}_X^{-1}(l-1)\mathbf{X}(l-1)\mathbf{X}^t(l-1)\widehat{\mathbf{K}}_X^{-1}(l-1)}{\alpha+\beta(l)}\right], \tag{4.20}$$

where $\beta(l) = \mathbf{X}^t(l-1)\widehat{\mathbf{K}}_X^{-1}(l-1)\mathbf{X}(l-1)$. For later convenience, define the vector $\mathbf{g}(l)$ to be a portion of this recursion

$$\mathbf{g}(l) = \frac{\widehat{\mathbf{K}}_X^{-1}(l-1)\mathbf{X}(l-1)}{\alpha+\beta(l)}$$

and we thus express the estimate's inverse by

$$\widehat{\mathbf{K}}_X^{-1}(l) = \frac{1}{\alpha}\left[\widehat{\mathbf{K}}_X^{-1}(l-1) - \mathbf{g}(l)\mathbf{X}^t(l-1)\widehat{\mathbf{K}}_X^{-1}(l-1)\right]$$

A similar recursive structure can be imposed on the second term in the expression for the *FIR* Wiener filter's unit-sample response that started all of this ($\mathbf{h}_\diamond = \mathbf{K}_X^{-1}\mathbf{k}_{S\widetilde{S}}$). The vector $\mathbf{k}_{S\widetilde{S}}$ arose from the cross-covariance of the to-be-estimated signal with the observations. This quantity can be approximated by the recursion

$$\widehat{\mathbf{k}}_{S\widetilde{S}}(l) = \alpha\widehat{\mathbf{k}}_{S\widetilde{S}}(l-1) + S(l)\mathbf{X}(l-1)$$

The filter coefficients $\mathbf{h}(l+1)$ are given by $\widehat{\mathbf{K}}_X^{-1}(l)\widehat{\mathbf{k}}_{S\widetilde{S}}(l)$, which on substitution of the recursions becomes

$$\mathbf{h}(l+1) = \mathbf{h}(l) + \mathbf{g}(l)\left[S(l) - \mathbf{h}^t(l)\mathbf{X}(l-1)\right]$$

*This requirement is needed because we are focused on prediction problems.

†Note that the frames used in both estimates overlap. If desired, nonoverlapping frames could be used with the advantage of a more statistically stable estimate. Observations would span a longer period, however.

The equations governing the *RLS* algorithm are therefore

$$\begin{aligned}
\widehat{S}(l) &= \mathbf{h}^t(l)\mathbf{X}(l-1) \\
\varepsilon(l) &= S(l) - \widehat{S}(l) \\
\beta(l) &= \mathbf{X}^t(l-1)\widehat{\mathbf{K}}_X^{-1}(l-1)\mathbf{X}(l-1) \\
\mathbf{g}(l) &= \frac{\widehat{\mathbf{K}}_X^{-1}(l-1)\mathbf{X}(l-1)}{\alpha+\beta(l)} \\
\widehat{\mathbf{K}}_X^{-1}(l) &= \frac{1}{\alpha}\left[\widehat{\mathbf{K}}_X^{-1}(l-1) - \mathbf{g}(l)\mathbf{X}^t(l-1)\widehat{\mathbf{K}}_X^{-1}(l-1)\right] \\
\mathbf{h}(l+1) &= \mathbf{h}(l) + \mathbf{g}(l)\varepsilon(l)
\end{aligned}$$

The difference equation for updating the inverse of the covariance matrix is numerically sensitive. This computational issue can be approached using so-called square-root algorithms that update the matrix square root of the covariance matrix [9]. The parameter α in this algorithm plays a similar role as μ in the *LMS* approach. Both parameters control the effective amount of data used in the averaging and hence the degree to which the algorithms can track variations in signal characteristics. The *RLS* algorithm *always* converges so long as $0 < \alpha < 1$; otherwise, the covariance estimate recursion is unstable. In addition, α controls the variance of the long-term estimate. The misadjustment factor of the *RLS* algorithm is approximately equal to $M_{\text{RLS}} = (1-\alpha)L$ [36]. Both procedures require initial conditions to begin their recursions. Usually, the filter coefficients are initially zero. The *RLS* algorithm requires an additional quantity: the covariance matrix of the observations. A typical choice for this initial covariance matrix estimate is a constant times the identity matrix, the constant equaling an estimate of the signal variance.

Example

Consider the first-order example given when we considered the Wiener filter {88} where no observation noise is present. The initial observation occurs at $l = 0$. Given that the statistical characteristics of the signal are known, the optimal one-step predictor is a constant times the previous observations. The adaptive algorithms, *LMS* and *RLS*, do not presume knowledge of signal characteristics *other* than filter duration. The optimal one-step predictor for a first-order signal is a zeroth-order system ($L = 1$). This type of consideration is an example of a structural constraint; usually, the duration of the adaptive filter is chosen longer than that expected to be required by the data. This choice must be made carefully as it directly influences the amount of computation in both algorithms. For simplicity, set $L = 1$.

The *LMS* algorithm requires a value for μ to ensure convergence; too large a choice leads to an unstable filter. In this simple, first-order example, the variance σ_X^2 of the observations is about 2.5 and μ must be less than 0.13.* The *RLS* algorithm can also use an estimate of the observations' variance to initialize the covariance matrix estimate. The choice of α is governed by the misadjustment factor. To compare fairly the two adaptive algorithms, we set the misadjustment factors equal. The values of the single filter coefficient are shown for each in Fig. 4.6 {96}. The *RLS* filter coefficient settles much more rapidly than the *LMS* one. The *LMS* algorithm must both "sense" the signal characteristics *and* search for the optimal filter; the *RLS* algorithm needs only to perform the latter chore.

The *RLS* algorithm, which generally produces better adaptive filters by most any measure, does require more computations than the *LMS* filter *in its present form*. Typically, the *RLS* algorithm converges quickly, taking about $2M$ steps to converge [45: 395]. This convergence rate is much greater than that provided by the *LMS* algorithm. The quickened convergence is bought with increased computational complexity: Computations for each *RLS* update are of order $O(L^2)$ and $O(L)$ for *LMS*. The computations of the covariance

*Because the system is first order, we use the constraint $\mu < 1/3\sigma_X^2$ (Eq. 4.19 {95}).

matrix's inverse can be reduced by exploiting the matrix's Toeplitz symmetries for both algorithms. Generically known as fast Kalman algorithms [3: 142–74];[45: 409–44];[67], they provide linear computational complexity for both algorithms.

4.4.4 Kalman Filters

Kalman filters are linear minimum mean-squared error waveform estimators that generalize Wiener filters to situations in which the observations have time-varying statistics [4];[37];[45: 269–301];[55]. On the surface, this problem may seem esoteric. Kalman filters, however, directly face the issue ducked by the Wiener filter: initiating a linear waveform estimator from the first observation. Wiener filters *tacitly* assume that observations have been forever available, an unrealistic requirement. More important, the structure of the Kalman filter provides direct insights into the nature of *MMSE* waveform estimators. For example, we should have noted previously that the Wiener filter's output does *not* have a power spectrum equal to that assumed for the signal. A simple illustration of this point is the noncausal Wiener filter; the power spectrum of its waveform estimate is

$$\mathscr{S}_{\hat{s}}(\omega) = \frac{\mathscr{S}_s^2(\omega)}{\mathscr{S}_s(\omega) + \mathscr{S}_N(\omega)}$$

This power spectrum tends toward the signal's power spectrum $\mathscr{S}_s(\omega)$ only when that of the noise tends to zero, not a very interesting situation. Why shouldn't the optimum filter make the power spectrum of its output equal that of the signal? This question and others are best answered in the context of the Kalman filter.

Rather than use the unit-sample response framework of linear waveform estimation developed in §4.4.1, Kalman filters are derived in terms of *state variable* characterizations of linear systems. This re-expression is, in fact, the key to gaining more insight into linear waveform estimators. The Kalman filter provides the basis for *tracking* systems that coalesce the outputs from several arrays to locate a moving source of propagating energy. Here, the concept of state becomes crucial to developing effective algorithms. Conversely, signals having zeros in their power spectra cannot be easily represented by state variable characterizations; the gain in insight more than compensates for the loss of generality.

State-variable modeling of observations. Signal production is modeled as passing white noise through a system characterized by the state vector $\mathbf{v}$. The *state equation*, which governs the behavior of the system, is given by

$$\mathbf{v}(l) = \mathbf{A}(l)\mathbf{v}(l-1) + \mathbf{u}(l) \tag{4.21}$$

The *observation equation* describes the signal $\mathbf{s}(l)$ as a linear combination of state vector components $\mathbf{C}(l)\mathbf{v}(l)$ observed in the presence of additive noise.

$$\mathbf{X}(l) = \underbrace{\mathbf{C}(l)\mathbf{v}(l)}_{\mathbf{s}(l)} + \mathbf{N}(l) \tag{4.22}$$

Note that the observations and the signal can be vectors in this characterization; in this way, multiple, interdependent observations can be described and their estimates described as the result of a single, unified operation. In previous sections, the signal and noise models have been constrained to stationary random sequences. The explicit dependence of every aspect of the state and observation equations on the index l permits relaxation of that constraint.

The usual questions you would ask about a characterization of a linear, time-invariant system — what are the transfer function and unit-sample responses — are easily answered. For the state equation (4.21) to describe a time-invariant system, the *state transition matrix* $\mathbf{A}(l)$ must not change with time. In this special case, we can easily calculate the transfer function by evaluating the z-transform of the state equation.

$$\begin{aligned} \mathbf{V}(z) &= \mathbf{A}\mathbf{V}(z)z^{-1} + \mathbf{U}(z) \\ \Longrightarrow \mathbf{V}(z) &= (\mathbf{I} - \mathbf{A}z^{-1})^{-1}\mathbf{U}(z) \end{aligned}$$

Therefore, the *matrix transfer function* equals $(\mathbf{I} - \mathbf{A}z^{-1})^{-1}$. This transfer function has only poles (no finite zeros). Its poles are found by solving $\det(\mathbf{I} - \mathbf{A}z^{-1}) = \det(z\mathbf{I} - \mathbf{A}) = 0$. This equation is identical to the one

needed to find the eigenvalues of $\mathbf{A}$.* Consequently, the poles of the transfer function equal the eigenvalues of $\mathbf{A}$. The corresponding eigenvectors are termed the *modes* of $\mathbf{A}$. To find the unit-sample response, we assume $\mathbf{v}(-1) = 0$ (otherwise, the system is not linear) and set $\mathbf{u}(l) = \boldsymbol{\delta}(l)$. Labeling the resulting output $\mathbf{h}(l)$, we have $\mathbf{h}(l) = \mathbf{A}^l\boldsymbol{\delta}(0)$, $l \geq 0$. As expected, the z-transform of this result equals the transfer function just derived. Interestingly, as shown in problem 4.36, the state matrix $\mathbf{A}$ and the choice of states $\mathbf{v}$ are not unique. However, the eigenvalues and eigenvectors remain the same, making these characteristic of the system, *not* of how it is described.

To show how a state description can be created from a single-input/single-output description, consider the difference equation $x(l) = a_1x(l-1) + a_2x(l-2) + u(l)$. The dimension of the state vector must equal the number of poles. One possible state vector is $\mathbf{v}(l) = \text{col}[x(l), x(l-1)]$. With this choice,

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 \\ 1 & 0 \end{bmatrix}$$

The observation matrix $\mathbf{C}$ would equal the row vector $[1\ 0]$.

Kalman filter derivation. In the Kalman filter scenario, the noise vectors are assumed to be zero-mean, white Gaussian random vectors.

$$\mathsf{E}[\mathbf{u}(k)\mathbf{u}^t(l)] = \mathbf{K}_u(l)\delta(k-l) \qquad \mathsf{E}[\mathbf{N}(k)\mathbf{N}^t(l)] = \mathbf{K}_N(l)\delta(k-l)$$

Note the *components* of these noise vectors are allowed to be correlated in this framework. Additive colored noise in the observations can be described by enlarging the number of states in the state equation and defining the matrix $\mathbf{C}$ appropriately.

The minimum mean-squared error *linear* estimate of the signal is equivalent to estimating the state vector and then multiplying by the observation matrix $\mathbf{C}$: $\widehat{\mathbf{S}}(l) = \mathbf{C}(l)\widehat{\mathbf{v}}(l)$. Therefore, we need to derive the optimal state estimate. The Kalman filter focuses on the *filtered* estimate: the estimate of the state at index l that depends on all of the observations made up to and including l, commencing with $\mathbf{X}(0)$. Extending this estimate to prediction problems is easy: $\widehat{\mathbf{S}}(l+l_e) = \mathbf{C}(l+l_e)\left[\prod_{k=l}^{l_e}\mathbf{A}(k)\right]\widehat{\mathbf{v}}(l)$, which equals $\mathbf{C}\mathbf{A}^{l_e}\widehat{\mathbf{v}}(l)$ in the time-invariant case.

Denote the state estimate by $\widehat{\mathbf{v}}(l|m)$, the notation intending to convey that the estimate at index l depends on *all* the observations made up to and including those made at index m $\{\mathbf{X}(m), \mathbf{X}(m-1), \ldots, \mathbf{X}(0)\}$. The filtered state estimate thus takes the form $\widehat{\mathbf{v}}(l|l)$; we assume it to have a state variable characterization:

$$\widehat{\mathbf{v}}(l|l) = \mathbf{M}(l)\widehat{\mathbf{v}}(l-1|l-1) + \mathbf{G}(l)\mathbf{X}(l)$$

This formulation is verified by finding the matrices $\mathbf{M}$ and $\mathbf{G}$ that permit the Orthogonality Principle to be obeyed: the estimation error $\boldsymbol{\varepsilon}(l|l) = \mathbf{v}(l) - \widehat{\mathbf{v}}(l|l)$ must be orthogonal to all linear transformations that can be applied to the observations. For instance, the waveform estimator is certainly a linear system acting on the observations; hence the estimator $\widehat{\mathbf{v}}(l-1|l-1)$ is a linear transformation of all the observations obtained before index l and must be orthogonal to each component of the estimation error.

$$\mathsf{E}\left[\{\mathbf{v}(l) - \widehat{\mathbf{v}}(l|l)\}\widehat{\mathbf{v}}^t(l-1|l-1)\right] = \mathbf{0}$$

This property can be used to find the matrix $\mathbf{M}$.†

$$\mathsf{E}\left[\{\mathbf{v}(l) - \mathbf{M}\widehat{\mathbf{v}}(l-1|l-1) - \mathbf{G}\mathbf{X}(l)\}\widehat{\mathbf{v}}^t(l-1|l-1)\right] = \mathbf{0}$$

or

$$\mathsf{E}\left[\{\mathbf{v}(l) - \mathbf{G}\mathbf{X}(l)\}\widehat{\mathbf{v}}^t(l-1|l-1)\right] = \mathsf{E}\left[\{\mathbf{M}\widehat{\mathbf{v}}(l-1|l-1)\}\widehat{\mathbf{v}}^t(l-1|l-1)\right]$$

Adding and subtracting the state $\mathbf{v}(l-1)$ from $\widehat{\mathbf{v}}(l-1|l-1)$ occurring in braces on the right side of this equation changes that term to $\mathbf{v}(l-1) - \boldsymbol{\varepsilon}(l-1|l-1)$. Applying the Orthogonality Principle results in

$$\mathsf{E}\left[\{\mathbf{v}(l) - \mathbf{G}\mathbf{X}(l)\}\widehat{\mathbf{v}}^t(l-1|l-1)\right] = \mathsf{E}\left[\{\mathbf{M}\mathbf{v}(l-1)\}\widehat{\mathbf{v}}^t(l-1|l-1)\right]$$

*See §B.5 for more on eigenanalysis.

†The dependence of the various matrices on time is notationally implicit in sequel. Consequently, reference to a matrix (such as $\mathbf{A}$) denotes the matrix $\mathbf{A}(l)$ at index l.

Using the observation equation (4.22) to eliminate $\mathbf{X}(l)$ leaves

$$\mathsf{E}\left[\{\mathbf{v}(l)-\mathbf{GCv}(l)-\mathbf{GN}(l)\}\widehat{\mathbf{v}}^t(l-1|l-1)\right]=\mathsf{E}\left[\{\mathbf{Mv}(l-1)\}\widehat{\mathbf{v}}^t(l-1|l-1)\right]$$

Because the noise $\mathbf{N}(l)$ is uncorrelated from index to index, it is uncorrelated with the state estimate $\widehat{\mathbf{v}}(l-1|l-1)$, which depends only on the noise values previous to index l. Thus, the expected value of the inner product between these two quantities is zero. Substituting the state equation (4.21) into the left side of this equation and noting a similar simplification with respect to the input noise $\mathbf{u}(l)$ leaves

$$\mathsf{E}\left[\left\{(\mathbf{I}-\mathbf{GC})\mathbf{Av}(l-1)\right\}\widehat{\mathbf{v}}^t(l-1|l-1)\right]=\mathsf{E}\left[\left\{\mathbf{Mv}(l-1)\right\}\widehat{\mathbf{v}}^t(l-1|l-1)\right]$$

or

$$\mathsf{E}\left[\{\mathbf{M}-(\mathbf{I}-\mathbf{GC})\mathbf{A}\}\mathbf{v}(l-1)\widehat{\mathbf{v}}^t(l-1|l-1)\right]=\mathbf{0}$$

For this equality to be satisfied for all nonzero states and their estimates, the matrix $\mathbf{M}$ must satisfy

$$\mathbf{M}=[\mathbf{I}-\mathbf{GC}]\mathbf{A}$$

The Kalman filter state estimator equation thus becomes

$$\boxed{\widehat{\mathbf{v}}(l|l)=\mathbf{A}(l)\widehat{\mathbf{v}}(l-1|l-1)+\mathbf{G}(l)\left[\mathbf{X}(l)-\mathbf{C}(l)\mathbf{A}(l)\widehat{\mathbf{v}}(l-1|l-1)\right]} \tag{4.23}$$

This rather complicated result hides an intriguing and useful interpretation. A critical aspect of this interpretation is the idea of the *one-step predictor* $\widehat{\mathbf{v}}(l|l-1)$: the estimate of the current state based on all previous observations. As no other observations are available, assume that the estimate is based solely on the state estimate at $l-1$: $\widehat{\mathbf{v}}(l-1|l-1)$. The predictor is therefore written

$$\widehat{\mathbf{v}}(l|l-1)=\mathbf{L}\widehat{\mathbf{v}}(l-1|l-1)$$

where $\mathbf{L}$ is a matrix found by satisfying the Orthogonality Principle.

$$\mathsf{E}\left[\{\mathbf{L}\widehat{\mathbf{v}}(l-1|l-1)-\mathbf{v}(l)\}\widehat{\mathbf{v}}^t(l-1|l-1))\right]=\mathbf{0}$$

or

$$\begin{aligned}
\mathsf{E}\left[\{\mathbf{L}\widehat{\mathbf{v}}(l-1|l-1)\}\widehat{\mathbf{v}}^t(l-1|l-1))\right] &= \mathsf{E}\left[\mathbf{v}(l)\widehat{\mathbf{v}}^t(l-1|l-1))\right] \\
&= \mathsf{E}\left[\{\mathbf{Av}(l-1)+\mathbf{u}(l)\}\widehat{\mathbf{v}}^t(l-1|l-1))\right] \text{ [use state equation]} \\
&= \mathsf{E}\left[\{\mathbf{Av}(l-1)\}\widehat{\mathbf{v}}^t(l-1|l-1))\right] \quad [\mathbf{u}(l) \text{ uncorrelated with past}] \\
&= \mathsf{E}\left[\{\mathbf{A}\widehat{\mathbf{v}}(l-1|l-1)\}\widehat{\mathbf{v}}^t(l-1|l-1))\right] \text{ [add/subtract } \widehat{\mathbf{v}}(l-1|l-1); \text{ use OP]}
\end{aligned}$$

Consequently, the matrix $\mathbf{L}$ equals $\mathbf{A}(l)$, which leaves the one-step predictor as

$$\boxed{\widehat{\mathbf{v}}(l|l-1)=\mathbf{A}(l)\widehat{\mathbf{v}}(l-1|l-1)}$$

In the Wiener filter example of estimating a first-order signal observed in noise {88}, we found the one-step predictor to be a scaled version of the previous estimate. We now see that the Kalman filter generalizes that finding to higher-order signals.

The estimator equation (4.23) can be more simply expressed in terms of the one-step predictor.

$$\boxed{\widehat{\mathbf{v}}(l|l)=\widehat{\mathbf{v}}(l|l-1)+\mathbf{G}[\mathbf{X}(l)-\mathbf{C}\widehat{\mathbf{v}}(l|l-1)]}$$

The term $\boldsymbol{\nu}(l)=\mathbf{X}(l)-\mathbf{C}\widehat{\mathbf{v}}(l|l-1)$ serves as the input and represents that part of the current observations $\mathbf{X}(l)$ *not* predicted by the previous ones. This quantity is often referred to as the *innovations sequence*. This basic form of the estimator equation occurs in many linear estimation situations: The value of the one-step predictor is augmented by an input term equal to a matrix $\mathbf{G}$, the *Kalman gain matrix*, multiplying the innovations sequence.

To derive the equations determining the Kalman gain, we need expressions for the estimation errors $\boldsymbol{\varepsilon}(l|l)$ and $\boldsymbol{\varepsilon}(l|l-1)$.

$$\begin{aligned}\boldsymbol{\varepsilon}(l|l-1) &= \mathbf{A}\boldsymbol{\varepsilon}(l-1|l-1)+\mathbf{u}(l)\\ \boldsymbol{\varepsilon}(l|l) &= \mathbf{A}\boldsymbol{\varepsilon}(l-1|l-1)+\mathbf{u}(l)-\mathbf{GCA}\boldsymbol{\varepsilon}(l-1|l-1)-\mathbf{G}[\mathbf{Cu}(l)+\mathbf{N}(l)]\\ \mathbf{X}(l) &= \mathbf{CA}[\boldsymbol{\varepsilon}(l-1|l-1)+\widehat{\mathbf{v}}(l-1|l-1)]+\mathbf{Cu}(l)+\mathbf{N}(l)\end{aligned}$$

Let $\mathbf{K}_\varepsilon$ denote the covariance matrix of the state estimation error: $\mathbf{K}_\varepsilon = \mathsf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t]$. The covariance matrix of the one-step prediction error is found from the first equation to be

$$\mathbf{K}_\varepsilon(l|l-1) = \mathbf{AK}_\varepsilon(l-1|l-1)\mathbf{A}^t+\mathbf{K}_u$$

Because of the Orthogonality Principle, $\mathsf{E}\left[\boldsymbol{\varepsilon}(l|l)\mathbf{X}^t(l)\right]=0$. Using the last two equations, this constraint yields

$$\mathbf{K}_\varepsilon(l|l-1)\mathbf{C}^t-\mathbf{G}[\mathbf{CK}_\varepsilon(l|l-1)\mathbf{C}^t+\mathbf{K}_N]=0$$

We find the Kalman gain matrix to be

$$\boxed{\mathbf{G}(l)=\mathbf{K}_\varepsilon(l|l-1)\mathbf{C}^t(l)\left[\mathbf{C}(l)\mathbf{K}_\varepsilon(l|l-1)\mathbf{C}^t(l)+\mathbf{K}_N(l)\right]^{-1}}$$

Expressing the covariance matrix of the estimation error directly, a simple recursion for it emerges.

$$\mathbf{K}_\varepsilon(l|l)=\mathbf{K}_\varepsilon(l|l-1)-\mathbf{G}(l)\mathbf{C}(l)\mathbf{K}_\varepsilon(l|l-1)$$

Thus, the Kalman gain matrix is found recursively from the covariance matrix of the estimation error and the covariance matrix of the one-step prediction error. This recursion amounts to a solution of the *Riccati equation*, the nonlinear difference equation governing the evolution of the gain matrix.

In summary, the Kalman filter equations are

Kalman Prediction Equations:

$$\boxed{\begin{aligned}\widehat{\mathbf{v}}(l|l) &= \mathbf{A}(l)\widehat{\mathbf{v}}(l-1|l-1)+\mathbf{G}(l)\left[\mathbf{X}(l)-\mathbf{C}(l)\mathbf{A}(l)\widehat{\mathbf{v}}(l-1|l-1)\right]\\ \widehat{\mathbf{S}}(l) &= \mathbf{C}(l)\widehat{\mathbf{v}}(l|l)\end{aligned}} \tag{4.24a}$$

Kalman Gain Equations:

$$\boxed{\begin{aligned}\mathbf{K}_\varepsilon(l|l-1) &= \mathbf{A}(l)\mathbf{K}_\varepsilon(l-1|l-1)\mathbf{A}^t(l)+\mathbf{K}_u(l)\\ \mathbf{G}(l) &= \mathbf{K}_\varepsilon(l|l-1)\mathbf{C}^t(l)\left[\mathbf{C}(l)\mathbf{K}_\varepsilon(l|l-1)\mathbf{C}^t(l)+\mathbf{K}_N(l)\right]^{-1}\\ \mathbf{K}_\varepsilon(l|l) &= \left[\mathbf{I}-\mathbf{G}(l)\mathbf{C}(l)\right]\mathbf{K}_\varepsilon(l|l-1)\end{aligned}} \tag{4.24b}$$

To perform these recursions, initial conditions must be specified. As observations are usually taken to begin at $l=0$, values for $\widehat{\mathbf{v}}(-1|-1)$ and $\mathbf{K}_\varepsilon(-1|-1)$ must somehow be established. The initial value $\widehat{\mathbf{v}}(-1|-1)$ equals the state estimate when no observations are available. Given no data, the minimum mean-squared error estimate of *any* quantity is the mean. As the noise terms in the state variable representation of the signal and the observations are assumed to have zero mean, the expected value of the state is 0, implying that the initial value of the state estimate should be 0. In this case, the variance of the estimation error equals that of the state. Thus, the initial condition of the Riccati equation solution is $\mathbf{K}_\varepsilon(-1|-1)=\mathsf{E}[\mathbf{v}(-1)\mathbf{v}^t(-1)]$.

Example

The simplest example of a Kalman filter waveform estimator occurs when the signal generation model is first-order. Recalling this type of example from the previous section {88}, its state variable characterization is

$$v(l)=\frac{1}{2}v(l-1)+w(l)\quad \mathbf{K}_w=[1]$$

and the observation equation is

$$X(l)=v(l)+N(l)\quad \mathbf{K}_N=[8/7]$$

Thus, the quantities that determine the solution of the Kalman filter components are $\mathbf{A} = [1/2]$ and $\mathbf{C} = [1]$. The estimator equation is, therefore,

$$\widehat{S}(l) = \frac{1}{2}\widehat{S}(l-1) + G(l)[X(l) - \frac{1}{2}\widehat{S}(l-1)]$$

The covariance matrices are simply variances and the equations determining the Kalman gain are scalar equations.

$$G(l) = \frac{\sigma_\varepsilon^2(l|l-1)}{\sigma_\varepsilon^2(l|l-1) + 8/7}$$
$$\sigma_\varepsilon^2(l|l-1) = \frac{1}{4}\sigma_\varepsilon^2(l-1|l-1) + 1$$
$$\sigma_\varepsilon^2(l|l) = \sigma_\varepsilon^2(l|l-1) - G(l)\sigma_\varepsilon^2(l|l-1)$$

The initial value for the variance, $\sigma_\varepsilon^2(-1|-1)$, equals the signal variance, which equals $4/3$. The table shows the values of the quantities involved in the Kalman gain computation.

l	$\sigma_\varepsilon^2(l\|l-1)$	$G(l)$	$\sigma_\varepsilon^2(l\|l)$
0	1.3333	0.5385	0.6154
1	1.1538	0.5024	0.5742
2	1.1435	0.5001	0.5716
3	1.1429	0.5000	0.5714
⋮	⋮	⋮	⋮
∞	8/7	1/2	4/7

Note how the gain varies despite the fact that the signal model is stationary. This effect occurs because the filter is started with no prior observation values. After few iterations that correspond to this initial transient, the Kalman filter coefficients settle and the signal estimate is governed by the difference equation

$$\begin{aligned}\widehat{S}(l) &= \frac{1}{2}\widehat{S}(l-1) + \frac{1}{2}[X(l) - \frac{1}{2}\widehat{S}(l-1)] \\ &= \frac{1}{4}\widehat{S}(l-1) + \frac{1}{2}X(l)\end{aligned}$$

which is precisely the Wiener filter result. The Kalman filter has the advantage that there is an explicit starting procedure; the Wiener filter is the Kalman filter in "steady state" for otherwise stationary signals.

4.5 Noise Suppression with Wavelets

Since their introduction in the mid 1980s, ideas from diverse scientific fields have resulted in developing wavelet signal analysis into a powerful analysis tool. The term *wavelet* is often used to denote a signal located in time with a concentrated amount of energy [14]. Wavelets are a set of basis functions that can represent an arbitrary function as a combination of scaled and dilated versions of a particular oscillatory "mother" function. The "mother" wavelet is used to generate a set of "daughter" functions through the operations of scaling and dilation applied to the mother wavelet. This set forms an orthogonal basis that allows, using inner products, to represent any given signal much like in the case of Fourier analysis. Wavelets, however, are superior to Fourier analysis in that both time and scale (analogous to frequency) information are represented in the transform domain.

The term *wavelet* was first introduced by Jean Morlet while working on the analysis of signals for seismic analysis on oil-related projects. Before Morlet's work remarkable contributions were developed by Haar [42]

and Zweig in 1975. After the work of Morlet and Grossmann on the definition of the continous wavelet transform (CWT) [40], several developings have followed. The work of researchers as Stromberg, Daubechies, Mallat and Newland, among others, has pushed forward the theoretical frontiers of wavelets-based orthogonal decomposition and also augmented the scope of possible application fields.

4.5.1 Wavelet Expansions

This review of wavelet signal representation is based on [14]. A wavelet is a known signal with special characteristics that allow it to be employed for studying the properties of other signals simultaneously in the frequency and time domains. A typical wavelet is shown in Figure 4.7. Based on a particular wavelet, it is possible to define a *wavelet expansion*. A wavelet expansion is the representation of a signal in terms of an orthogonal collection of real-valued functions generated by applying suitable transformations to the original wavelet. These functions are called "daughter" wavelets while the original wavelet is dubbed "mother" wavelet, acknowledging its function as source of the orthogonal collection. If $x(t)$ is a given signal to be decomposed, the signal can be represented by an orthonormal expansion as

$$x(t) = \sum_i a_i \psi_i(t) \ . \tag{4.25}$$

Here, $\{\psi_i(t)\}$ comprise a set of orthonormal basis functions. The coefficients a_i can be found through the inner product of $f(t)$ and the functions $\psi_i(t)$.

$$a_i = \langle x(t), \psi_i(t) \rangle = \int x(t)\psi_i(t)dt$$

In the special case of a wavelet expansion, the wavelet basis functions have two integer indexes as they are generated from the so-called "mother" wavelet by *scaling* and *translation* operations.

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \quad j,k \in \mathbb{Z} \tag{4.26}$$

and (4.25) must be rewritten as

$$x(t) = \sum_j \sum_k a_{j,k}\psi_{j,k}(t) \tag{4.27}$$

Equation (4.27) is termed the *wavelet expansion* of $x(t)$ while the collection of coefficients $a_{j,k}$ form the discrete wavelet transform (DWT) of $f(t)$.

Most wavelets basis functions satisfy *multi-resolution* conditions. This property guarantees that if a set of signals can be represented by basis functions generated from a translation $\psi(t-k)$ of the mother wavelet, then a larger set of functions, including the original, can be represented by a new set of basis functions $\psi(2t-k)$. This feature is used in the *fast wavelet transform* (FWT) algorithm, the equivalent of the FFT algorithm for wavelet decomposition. The computational complexity of the discrete wavelet transform is at most $O(L\log L)$, akin to the complexity of using the Fast Fourier Transform to calculate the discrete Fourier transform (DFT). For some particular types of wavelets, the complexity can be as low as $O(L)$.

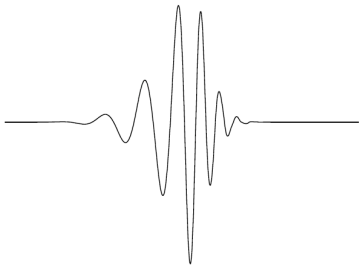


Figure 4.7: Daubechies wavelet $\psi D20$ [14].

Perhaps the most important feature of the wavelet expansion is that it provides a dual time-frequency localization of the input signal. Here, scaling — the contraction of the mother wavelet — amounts to selecting higher frequency components. The time shifts localize the analysis to specific time intervals when the mother wavelet has finite duration. This combination of properties implies that most of the energy of the signal will be captured by a few wavelet coefficients. The lower resolution coefficients can be calculated from the higher resolution coefficients using a *filter bank* algorithm [14]. This property contributes to the efficiency of the calculation of the FWT.

4.5.2 Denoising with Wavelets

Denoising means removing noise — unwanted information — that contaminates observations of an unknown signal. The use of wavelets for noise removal was first introduced by Donoho and Johnstone [30]. The shrinkage methods for noise removal, first introduced by Donoho [29], have led to a variety of approaches to signal denoising. If $X(t)$ is an empirically recorded (sampled) signal consisting of a signal $s(t)$ and additive noise.

$$X(t_l) = s(t_l) + \sigma N(t_l), \quad l = 1, \ldots, L \tag{4.28}$$

where $N(t_l)$ are independent normal random variables — $N(t_l) \sim \mathcal{N}(0,1)$ — and σ represents the intensity (standard deviation) of the noise in the observations $X(t_l)$. Given a finite set of $X(t_l)$ values, denoising attempts to reconstruct the original signal s without assuming a particular structure for it. The usual approach to noise removal assumes some knowledge of the signal's structure. In the case of Wiener filters, the signal is assumed to be described as a stochastic process having a known power spectrum. For Kalman filters, the characteristics of the signal's dynamic model must be known. The wavelet approach decomposes the observations using a "dual" frequency-time representation. Under this approach, noise will be represented as a white process having a constant power spectrum while the signal has a non-flat spectrum. Consequently, the wavelet coefficients of the noise are identically distributed Gaussian random variables.

Once the noise characteristics — its variance in particular — are identified, the removal process starts. It has been shown that a suitable strategy for noise removal consists in making the expansion coefficients associated to noise frequencies (scales) equal to zero, a process known as *thresholding*. Because this approach can be nonlinear, we avoid using the term "filtering," using the more general term "denoising" instead. This approach represents a global approach for noise removal, regardless of whether wavelets serve as the basis functions or not. The various methods for denoising differ in the way the signal coefficients are tracked and the noise coefficients removed from the representation.

Before attempting to describe the methods is convenient to discuss an alternative definition for wavelet representation used for noise removal. First, the description assumes that the representation is achieved using periodized wavelet bases on $[0,1]$. Also, the basis functions are generated by dilation and translation of a compactly supported (finite duration) scaling function ϕ, also called *father wavelet* and the familiar mother wavelet function ψ. ψ must be associated with an r-regular (i.e., the signals are smooth) multiresolution analysis of $L^2(\mathbb{R})$. An advantage of this approach is that generated wavelet families allow integration of different kinds of smoothness and vanishing moments. This feature leads to the fact that many signals in practice can be represented sparsely (with few wavelets coefficients) and uniquely with a wavelet decomposition. Using father and mother wavelets in a signal expansion,

$$s(t) = \sum_{k=0}^{2^{j_0}-1} c^s_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d^s_{j,k}\psi_{j,k}(t), \quad j_0 \geq 0, \quad t \in [0,1]$$

where j_0 is a primary resolution level, and $c^s_{j_0,k}$ and d^s_{jk} are calculated with inner products with the signal.

$$c^s_{j_0,k} = \langle s, \phi_{j_0,k}\rangle = \int_0^1 s(t)\phi_{j_0,k}(t)dt,\ j_0 \geq 0, \quad k = 0,1,\ldots,2^{j_0}-1$$

$$d^s_{j,k} = \langle s, \psi_{j,k}\rangle = \int_0^1 s(t)\psi_{j,k}(t)dt,\ j \geq j_0 \geq 0, \quad k = 0,1,\ldots,2^{j}-1$$

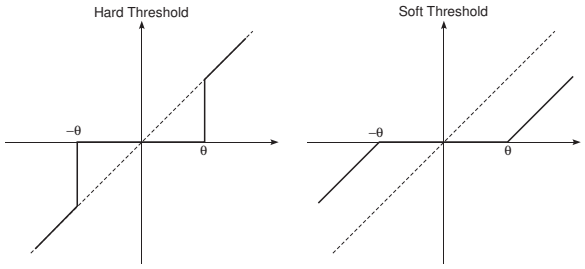


Figure 4.8: The "hard" threshold function is shown on the left, the "soft" threshold function on the right. θ represents the threshold value, which must be estimated from the data.

Applying the DWT to (4.28), the wavelet coefficients for the observations are

$$c^x_{j_0,k} = c^s_{j_0,k} + \sigma c^n_{j_0,k}, \quad k = 0,1,\ldots,2^{j_0}-1$$
$$d^x_{j,k} = d^s_{j,k} + \sigma d^n_{j,k}, \quad j = j_0,\ldots,J-1, \quad k = 0,\ldots,2^j-1$$

Classical approach to wavelet thresholding

The simplest way to remove noise from a contaminated signal consists in modifying the observation's wavelet coefficients in a smart way so that the "small" coefficients associated to the noise are neglected. The updated coefficients can thus be used to reconstruct the original underlying function free from the effects of noise. Implicit in this strategy is the presumption that only a few "large" wavelets coefficients $d^x_{j,k}$ are associated with the original signal, and that their identification and subsequent elimination of the other coefficients together allows perfect reconstruction of the underlying signal s. Several methods use this idea and implement it in different ways.

In the linear penalization method, every wavelet coefficient is affected by a linear *shrinkage* particular to the resolution level of the coefficient.

$$\tilde{d}^x_{jk} = \frac{d^x_{jk}}{1+\lambda 2^{2jr}}$$

Here, the parameter r is the *known* smoothness index of the underlying signal s, while parameter λ is a smoothing factor whose determination is critical for this type of analysis. It must be said that linear thresholding is adequate only for a stationary signal with specific levels of regularity. When homegeneity and regularity conditions are not met, nonlinear wavelet thresholding or shrinkage methods are usually more suitable.

Donoho et. al [28, 30] proposed a nonlinear strategy for thresholding. With their approach, the thresholding is implemented with either a hard or a soft thresholding rule (see Figure 4.8).

$$\tilde{d}^x_{j,k} = \begin{cases} 0 & |d^x_{j,k}| \le \theta \\ d^x_{j,k} & |d^x_{j,k}| > \theta \end{cases} \qquad \text{(hard)}$$

$$\tilde{d}^x_{j,k} = \begin{cases} 0 & |d^x_{j,k}| \le \theta \\ d^x_{j,k} - \theta & d^x_{j,k} > \theta \\ d^x_{j,k} + \theta & d^x_{j,k} < -\theta \end{cases} \qquad \text{(soft)}$$

In both methods, the role of the parameter θ as a threshold value is critical to the destruction, reduction or maintenance of a wavelet coefficient's value. Several authors have discussed the properties and limitations of these two strategies. Hard thresholding, due to its induced discontinuity, can be unstable and sensitive to small changes in the data. On the other hand, soft thresholding can create unnecessary bias when the true coefficients are large. Although more sophisticated methods has been introduced to account for the drawbacks of the described nonlinear strategies, the discussion here is limited to the hard and soft approaches.

Term-by-Term Thresholding

One apparent problem in applying wavelet thresholding methods is the way of selecting an appropriate value for the threshold θ. Strategies for determining the threshold can be classified in two groups: *global* thresholds and *level-dependent* thresholds. Global threshold implies the selection of one threshold value applied to all the wavelet coefficients. Level-dependent thresholds implies that a (possibly) different threshold value θ_j is applied at each scale. All the alternatives require an estimate of the noise level σ. The standard deviation of the data values is clearly not a good estimator because of the presence of the signal. Donoho [30] proposed estimating σ in the wavelet domain by using

$$\hat{\sigma} = \frac{\text{median}(|d^x_{J-1,k}|)}{0.6745},\ k = 0, 1, \ldots, 2^{J-1} - 1$$

Minimax threshold. Donoho [30] obtained an optimal threshold value θ^M by minimizing the risk (error) involved in estimating the signal. The proposed *minimax* threshold depends of the available data and also takes into account the noise level contaminating the signal.

$$\theta^M = \hat{\sigma}\theta^*$$

θ^* is equal to the value of θ satisfying

$$\theta^* = \inf_{\theta} \sup_{d} \left\{ \frac{R_\theta(d)}{L^{-1} + R_{\text{oracle}}(d)} \right\}$$

$R_\lambda(d)$ is calculated according to

$$R_\lambda(d) = \mathsf{E}\left[\left(\delta\left[d^2\right]\right)\right]$$

while $R_{\text{oracle}}(d)$ accounts for the risk associated to the modification of the value of a given wavelet coefficient and $\delta[\cdot]$ is the shrinkage/thresholding operation. Two oracles were introduced by Donoho [30]: *diagonal linear projection* (DLP) and *diagonal linear shrinker* (DLS).

$$R^{DLP}_{\text{oracle}}(d) = \min\{d^2, 1\}$$

$$R^{DLS}_{\text{oracle}}(d) = \frac{d^2}{d^2 + 1}$$

Antoniadis [5] provides values of the minimax threshold for both the hard and soft nonlinear thresholding rules. For the soft rule, 1.669 and 2.226 for L equal to 128 and 1024; for the hard rule, 2.913 and 3.497 for the same respective number of observations.

Universal threshold. Donoho [30] proposed this threshold as an alternative to the minimax thresholds that could be applied to all the wavelet coefficients.

$$\theta^U = \hat{\sigma}\sqrt{2\log L}$$

This threshold is easy to calculate and the optimization implicit in the minimax method is avoided. The universal threshold ensures, with high probability, that every value in the wavelet transform for which the underlying function is exactly zero will be estimated as zero. Note that the resulting convergance rate (depending in the size of the sample) can be slow.

How well denoising works is illustrated in Figure 4.9. Hard thresholding tends to better preserve the signal at the expense of allowing extreme noisy values to appear while soft thresholding gives a smoother result but one that departs significantly from the actual signal value. This scaling is due to the reduction of the values of wavelet coefficients even if they are only related to the signal.

Classical Methods: Block Thresholding

Thresholding approaches resorting to term-by-term modification of the wavelet coefficients attempt to balance variance and bias contributions to the mean squared error incurred in estimating the underlying signal s. However, it has been proven that such balance is not optimal. Term-by-term thresholding removes to many terms, leading to an estimate prone to bias and with a slower convergance rate due to the number of

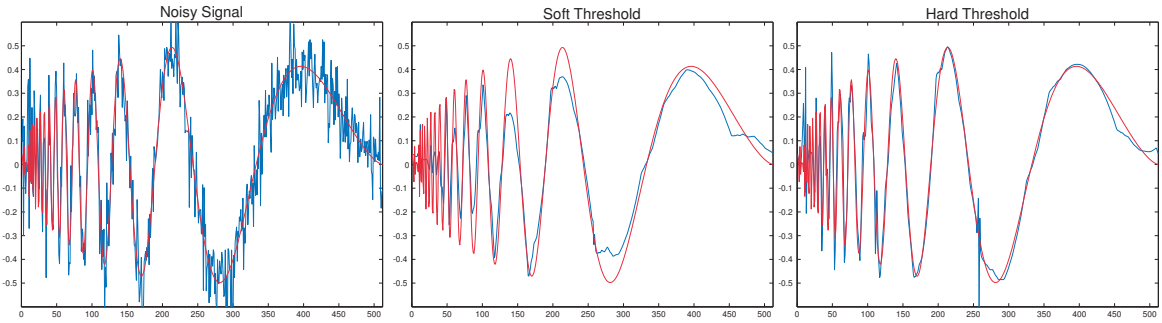


Figure 4.9: The original chirp signal is shown in red. The noisy signal to be denoised is shown in the left panel. Results from soft and hard thresholding are shown in the center and right panels respectively.

operations involved. The quality of wavelet denoising can be improved by using information of the entire set of data associated to a particular wavelet coefficient. To do so, block strategies for thresholding have been proposed [16, 48]. The main idea consists of isolating a block of wavelet coefficients and, based upon the information collected about the entire set, make a decision about decreasing or even entirely discard the group. This procedure allows faster manipulation of the information and accelerated convergence rates.

4.6 Particle Filtering

The typical state-space formulation of a linear (possibly time-varying) system in discrete time is defined by

$$\text{Linear State Equation:} \quad \mathbf{v}(l) = \mathbf{A}\mathbf{v}(l-1) + \mathbf{B}\mathbf{u}(l) \tag{4.29}$$

$$\text{Linear Observation Equation:} \quad \mathbf{y}(l) = \mathbf{C}\mathbf{v}(l) + \mathbf{w}(l) \tag{4.30}$$

Here, the system's inputs are the elements of $\mathbf{u}(l)$ and the outputs are the elements of $\mathbf{y}(l)$. The vector $\mathbf{v}(l)$ denotes the *state vector*; its dimension determines the system's *order* M (how many poles it has) and thus its complexity. The *state matrix* $\mathbf{A}$ is a $M \times M$ square matrix that determines the system's dynamics. The output matrix $\mathbf{C}$ determines what linear combination of the state vector's elements contribute to the output. Added to this combination is observational noise $\mathbf{w}(l)$. Note that the state equation (4.29) entirely determines the system's dynamics. However, the state equation and the output matrix are *not* unique for a fixed input-output relationship between the input and output. Usually, convenient forms for the state (and thus the state matrix) are used to model a given system.

In nonlinear cases, the same basic modeling idea is used.

$$\text{Nonlinear State Equation:} \quad \mathbf{v}(l) = \mathbf{f}\big(\mathbf{v}(l-1), \mathbf{u}(l)\big) \tag{4.31}$$

$$\text{Nonlinear Observation Equation:} \quad \mathbf{y}(l) = \mathbf{g}\big(\mathbf{v}(l), \mathbf{w}(l)\big) \tag{4.32}$$

Again, the state equation defines the nonlinear nature of the *Markov* evolution of the state and the observation equation shows the output depends nonlinearly on the state's instantaneous value and the observation noise. Said another way, whether linear or nonlinear, the observations $\mathbf{y}(l)$ depend on the state vector $\mathbf{v}(l)$ that is *hidden* from direct observation.

The state-space model for both linear and nonlinear system amounts to a *hidden Markov model* (HMM) for the observations. The signal processing goal of many applications is to estimate from observations the unmeasured state and/or the input(s) $\mathbf{u}(l)$.* The minimal mean-squared error solution is *always* the conditional mean. Let the vector $\mathbf{h}(l)$ contain the hidden variables of interest (the components of the state and/or input vectors). Its optimal estimate based only on the observations made at the same time as the estimate is $\widehat{\mathbf{h}}(l) = \mathsf{E}[\mathbf{h}(l) \mid \mathbf{y}(l)]$. In all conceivable cases, better estimates (smaller mean squared-error) can be obtain by including more data (observations) into the estimate. These various estimators differ only in what

*In the cases where the state is to be estimated, a pre-defined state characterization (a matrix $\mathbf{A}$) is used so that a unique answer exists.

portion of the observations are used in the estimation. We use the notation $\mathbf{y}(l_1 : l_2)$ to denote the time series $\mathbf{y}(l_1), \mathbf{y}(l_1+1), \ldots, \mathbf{y}(l_2)$. Furthermore, we may well want to revise our estimate of previously estimated hidden variables as more observations are made. Consequently, we generally seek an estimate of $\mathbf{h}(0 : l)$.

Only in the simplest case—Gaussian inputs into a linear system with Gaussian observation noise—is the optimal (in the mean-squared error sense) estimate known. It can be expressed either as the Kalman or the Wiener filter. If the state equation is nonlinear or the input (in particular) is non-Gaussian, these optimal solutions for the linear and Gaussian case generally become suboptimal. To find the optimal estimate, we need the conditional probability distribution $p\big(\mathbf{h}(0 : l) \mid \mathbf{y}(l_1 : l_2)\big)$; in particular, we focus on $p\big(\mathbf{h}(0 : l) \mid \mathbf{y}(0 : l)\big)$. We therefore seek $p\big(\mathbf{v}(0 : l) \mid \mathbf{y}(0 : l)\big)$. Note that this conditional probability distribution is "backwards:" it is the probability density of the input given the output. The problem statement (equations 4.29 and 4.30 or equations 4.31 and 4.32) defines the "forward" distribution: the conditional distribution of the output given the input.

4.6.1 Recursive Framework

All filtering algorithms exploit the special structure of the state-space formalism to reduce computations. In particular, they use the relationship between $p\big(\mathbf{h}(l) \mid \mathbf{y}(0 : l)\big)$ and $p\big(\mathbf{h}(l-1) \mid \mathbf{y}(0 : l-1)\big)$ to update the conditional distribution on an index-by-index basis. First of all, we note that the state and observations obey the Markovian property

$$\begin{aligned} p\big(\mathbf{v}(l-1), \mathbf{v}(l) \mid \mathbf{y}(0 : l-1)\big) &= p\big(\mathbf{v}(l) \mid \mathbf{v}(l-1), \mathbf{y}(0 : l-1)\big)\, p\big(\mathbf{v}(l-1) \mid \mathbf{y}(0 : l-1)\big) \\ &= p\big(\mathbf{v}(l) \mid \mathbf{v}(l-1)\big)\, p\big(\mathbf{v}(l-1) \mid \mathbf{y}(0 : l-1)\big) \end{aligned}$$

The simplification in going from the first to the second amounts to the Markov property of the state equations (4.29) and (4.31): the previous state and the input completely determine the statistical properties of the current state. Integrating out $\mathbf{v}(l-1)$ we find that

$$p\big(\mathbf{v}(l) \mid \mathbf{y}(0 : l-1)\big) = \int p\big(\mathbf{v}(l) \mid \mathbf{v}(l-1)\big)\, p\big(\mathbf{v}(l-1) \mid \mathbf{y}(0 : l-1)\big)\, d\mathbf{v}(l-1) \tag{4.33}$$

This equation is known as the *Chapman-Kolmogorov equation* and occurs frequently in Markovian problems. If we augment the state vector with components of the input $\mathbf{u}(l)$ that form a white noise sequence (but which are statistically independent rather than correlated), the same result applies.

$$p\big(\mathbf{h}(l) \mid \mathbf{y}(0 : l-1)\big) = \int p\big(\mathbf{h}(l) \mid \mathbf{h}(l-1)\big)\, p\big(\mathbf{h}(l-1) \mid \mathbf{y}(0 : l-1)\big)\, d\mathbf{h}(l-1)$$

To find the conditional distribution we seek, we use Bayes' Rule.

$$\begin{aligned} p\big(\mathbf{h}(l) \mid \mathbf{y}(0 : l)\big) &= p\big(\mathbf{h}(l) \mid \mathbf{y}(0 : l-1), \mathbf{y}_n\big) && \text{(a)} \\ &= \frac{p\big(\mathbf{y}_n \mid \mathbf{h}_n, \mathbf{y}(0 : l-1)\big)\, p\big(\mathbf{h}_n \mid \mathbf{y}(0 : l-1)\big)}{p\big(\mathbf{y}_n \mid \mathbf{y}(0 : l-1)\big)} && \text{(b)} \\ &= \frac{p\big(\mathbf{y}(l) \mid \mathbf{h}(l)\big)\, p\big(\mathbf{h}(l) \mid \mathbf{y}(0 : l-1)\big)}{p\big(\mathbf{y}(l) \mid \mathbf{y}(0 : l-1)\big)} && (4.34) \end{aligned}$$

Equation (a) simply expands the observations into previous ones and the current one, equation (b) is Bayes' Rule and (4.34) uses the Markov property again. Consequently, (4.33) and (4.34) provide the exact way of updating the conditional distribution $p\big(\mathbf{h}(l-1) \mid \mathbf{y}(0 : l-1)\big)$ to find $p\big(\mathbf{h}(l) \mid \mathbf{y}(0 : l)\big)$. The update requires knowledge of the state-space equation to find $p\big(\mathbf{h}(l) \mid \mathbf{h}(l-1)\big)$ and the observation equation to find $p\big(\mathbf{y}(l) \mid \mathbf{h}_n\big)$.* The conditional distribution $p\big(\mathbf{y}(l) \mid \mathbf{y}(0 : l-1)\big)$ amounts to normalization: it equals the integral of the numerator of (4.34) with respect to $\mathbf{h}(l)$.

*In the latter equation, conditioning on the state suffices as the observation equation depends only on the state and not on the input.

A similar derivation can be used to derive how to update $p\big(\mathbf{h}(0:l-1) \mid \mathbf{y}(0:l-1)\big)$ to obtain $p\big(\mathbf{h}(0:l) \mid \mathbf{y}(0:l)\big)$.

$$\begin{aligned} p\big(\mathbf{h}(0:l) \mid \mathbf{y}(0:l)\big) &= \frac{p\big(\mathbf{y}(l) \mid \mathbf{y}(0:l-1), \mathbf{h}(0:l)\big)\, p\big(\mathbf{h}(0:l) \mid \mathbf{y}(0:l-1)\big)\, \cancel{p\big(\mathbf{y}(0:l-1)\big)}}{p\big(\mathbf{y}_n \mid \mathbf{y}(0:l-1)\big)\, \cancel{p\big(\mathbf{y}(0:l-1)\big)}} \\ &= \frac{p\big(\mathbf{y}(l) \mid \mathbf{y}(0:l-1), \mathbf{h}(0:l)\big)\, p\big(\mathbf{h}(l) \mid \mathbf{h}(0:l-1)\mathbf{y}(0:l-1)\big)\, p\big(\mathbf{h}(0:l-1) \mid \mathbf{y}(0:l-1)\big)}{p\big(\mathbf{y}_n \mid \mathbf{y}(0:l-1)\big)} \\ &= \frac{p\big(\mathbf{y}(l) \mid \mathbf{h}(l)\big)\, p\big(\mathbf{h}(l) \mid \mathbf{h}(l-1)\big)\, p\big(\mathbf{h}(0:l-1) \mid \mathbf{y}(0:l-1)\big)}{p\big(\mathbf{y}_n \mid \mathbf{y}(0:l-1)\big)} \\ &\propto p\big(\mathbf{y}(l) \mid \mathbf{h}(l)\big)\, p\big(\mathbf{h}(l) \mid \mathbf{h}(l-1)\big) \cdot p\big(\mathbf{h}(0:l-1) \mid \mathbf{y}(0:l-1)\big) \end{aligned} \tag{4.35}$$

The last step relies on the Markov property of the state equation and the nature of the observation equation.

When the input and noise terms are Gaussian and the linear model of (4.29) and (4.30) applies, the component probability distributions are Gaussian, making the result Gaussian. In this case, finding the conditional mean required for the estimate is straightforward. Performing these calculations results in both the Kalman and Wiener filters. In virtually all other cases, calculating the Chapman-Kolmogorov equation (4.33) is impossible, leading us to seek numerical solutions. Computationally more efficient are Monte Carlo simulations that use statistical methods to estimate the required distributions and evolve them recursively. Particle filters are one of these methods.

4.6.2 Estimating Probability Distributions using Monte Carlo Methods

Before we tackle the problem at hand, consider first how to generate random variables that have a specific distribution. The most common technique is based on the *cumulative probability distribution* function $P_Z(z)$. It is easily shown that if you consider the probability distribution function as a function and map the random variable Z with it, you obtain a uniform random variable defined over $[0,1]$. Consequently, if you have a uniform random variable, $P_Z^{-1}(U)$ creates the original random variable. Consequently, given a pseudo-random generator that produces uniformly distributed values over the unit interval, you can generate simulated values of the random variable Z. Assuming $z^{(k)}, k = 1, \ldots, K$ are generated this way,* an empirical estimate of the probability distribution $p_Z(z)$ is given by

$$\widehat{p}_Z(z) = \frac{1}{K} \sum_{k=1}^{K} \delta(z - z^{(k)}) \,,$$

where $\delta(\cdot)$ is the impulse. In the parlance of particle filtering, the values $z^{(k)}$ are particles, each of which have the same weight $1/K$. This approximation for the probability density of a continuous random variable, like a Gaussian, would not seem to be very good. However, if all we are interested in is the mean, it can be quite good. It will turn out other statistical measures can also be derived from this crude estimate. Conceptually, the underlying probability distribution could be a conditional one. The generation process would simply need to incorporate the conditioning values. Looking ahead a little, since we want to update a probability distribution, we need to "migrate" the particles. We won't change their locations; we'll change their weights.

However, our method has a flaw: we need the probability distribution we seek to generate the particle locations. Furthermore, even if we did, there is no guarantee that generating them will be easy: the inverse of the probability distribution function may be intractable, for example. The statistical technique of *importance sampling* comes to the fore. Suppose it is easy to generate random variables having the probability distribution $q\big(\cdot\big)$, but we want random variables distributed according to $p(\cdot)$. Assuming these two distributions have the same support, the values $z^{(k)}, k = 1, \ldots, K$ generated according to $q\big(\cdot\big)$ can be used to estimate $p(\cdot)$ with

$$\widehat{p}(z) = \sum_{k=1}^{K} w^{(k)} \delta(z - z^{(k)}), \quad w^{(k)} \propto \frac{p\big(z^{(k)}\big)}{q\big(z^{(k)}\big)} \tag{4.36}$$

*We adopt the convention that the subscript expresses the time index range and the superscript indexes the simulated value.

The constant of proportionality for the weights is required so that $\sum_k w^{(k)} = 1$: we require the estimate to integrate to one. Do note, however, that importance sampling introduces variability into the estimate that would have not be present if the actual probability distribution had been used to generate the particles.

So far, we have applied importance sampling to the case of approximating the distribution of a single random variable. The same approach can be used for random vectors and for conditional distributions.

$$p\big(\mathbf{h}(0:l) \mid \mathbf{y}(0:l)\big) \approx \sum_k w_n^{(k)} \delta\big(\mathbf{h}(0:l) - \mathbf{h}(0:l)^{(k)}\big)$$

Here, the random variables are generated according to $q\big(\mathbf{h}(0:l) \mid \mathbf{y}(0:l)\big)$ and the weights are proportional to $p\big(\mathbf{h}(0:l) \mid \mathbf{y}(0:l)\big)/q\big(\mathbf{h}(0:l) \mid \mathbf{y}(0:l)\big)$. Also note that now the weights depend on the index l: we will need to update them recursively. Furthermore, we want to generate particles for $[0:l]$ by augmenting the particles for $[0:l-1]$ with new particles for the l^{th} index, rather than generating particles for the entire time frame. Consequently, we seek recursive formulations based on the properties of the equations governing the observations and the state.

The importance-sampling distribution is arbitrary; at this point, all we require is that it be easily generated. In fact, a distribution that does not depend on the measurements or the estimate of hidden variables could be used. A convenient choice is to force the importance sampling distribution to factor.

$$q\big(\mathbf{h}(0:l) \mid \mathbf{y}(0:l)\big) = q\big(\mathbf{h}(l) \mid \mathbf{h}(0:l-1), \mathbf{y}(0:l)\big) \cdot q\big(\mathbf{h}(0:l-1) \mid \mathbf{y}(0:l-1)\big)$$

This form suggests a way of updating the importance sampling distribution recursively and, just as importantly, a way of updating the weights in a similar fashion. Exploiting (4.35) and the assumed factorization of the importance sampling distribution, we have

$$\begin{aligned} w(l)^{(k)} &\propto \frac{p\big(\mathbf{y}(l) \mid \mathbf{h}^{(k)}(l)\big)\, p\big(\mathbf{h}^{(k)}(l) \mid \mathbf{h}^{(k)}(l-1)\big) \cdot p\big(\mathbf{h}^{(k)}(0:l-1) \mid \mathbf{y}(0:l-1)\big)}{q\big(\mathbf{h}^{(k)}(l) \mid \mathbf{h}^{(k)}(0:l-1), \mathbf{y}(0:l)\big) \cdot q\big(\mathbf{h}^{(k)}(0:l-1) \mid \mathbf{y}(0:l-1)\big)} \\ &= \frac{p\big(\mathbf{y}(l) \mid \mathbf{h}^{(k)}(l)\big)\, p\big(\mathbf{h}^{(k)}(l) \mid \mathbf{h}^{(k)}(l-1)\big)}{q\big(\mathbf{h}^{(k)}(l) \mid \mathbf{h}^{(k)}(0:l-1), \mathbf{y}(0:l)\big)} w(l-1)^{(k)} \end{aligned} \tag{4.37}$$

If we further assume that this distribution depends only on the last available observations and state, so that $q\big(\mathbf{h}(l) \mid \mathbf{h}(0:l-1), \mathbf{y}(0:l)\big) = q\big(\mathbf{h}(l) \mid \mathbf{h}(l-1), \mathbf{y}(l)\big)$, the weight computation and simulation further simplify. This yields the weight computation

$$w(l)^{(k)} \propto \frac{p\big(\mathbf{y}(l) \mid \mathbf{h}^{(k)}(l)\big)\, p\big(\mathbf{h}^{(k)}(l) \mid \mathbf{h}^{(k)}(l-1)\big)}{q\big(\mathbf{h}^{(k)}(l) \mid \mathbf{h}^{(k)}(l-1), \mathbf{y}(l)\big)} w(l-1)^{(k)} \,. \tag{4.38}$$

This way, only the current values of the particles need be stored; past particles $\mathbf{h}^{(k)}(0:l-1)$ and past observations $\mathbf{y}(0:l-1)$ can be discarded. In the particle filtering "business," this choice for the importance sampling distribution is known as the *one-observation-ahead sampler*. This distribution is easy to calculate analytically (so that pseudo-random numbers can be generated) when the observation equation is linear as in (4.30).

Is this choice of importance sampling distribution a good one? To provide some measure of quality, we create a criterion that the choice should optimize. Most researchers focus on K_{eff}, the effective number of importance samples needed to achieve the same degree of simulation accuracy that using the true distribution would yield, and it is defined as

$$K_{\text{eff}} = \frac{K}{1 + \text{var}\left[w^{*(k)}(l)\right]} \,,$$

where $w^{*(k)}(l) = p\big(\mathbf{h}^{(k)}(l) \mid \mathbf{y}(0:l)\big)/q\big(\mathbf{h}^{(k)}(l) \mid \mathbf{h}^{(k)}(l-1), \mathbf{y}(l)\big)$ is known as the "true" weight. While it cannot be evaluated analytically, an estimate of it is

$$\widehat{K_{\text{eff}}} = \frac{1}{\sum_{k=1}^{K} \big(w(l)^{(k)}\big)^2} \,,$$

where the weights in this expression are the normalized weights calculated according to (4.37). Because each weight is less than one and squared, we known that $K_{\text{eff}} \leq K$. A small value of K_{eff} denotes *degeneracy* in the weights, which means that more than a few are small. Small weights contribute little to the probability density estimate as illustrated in (4.36). This degeneracy has been shown to occur always in particle filtering as the weights go through a succession of updates. If K_{eff} is too small, many particles have very small weights and accuracy in estimating even the mean, much less the probability distribution, is lost. The choice of importance sampling distribution that maximizes K_{eff} is

$$\begin{aligned} q\big(\mathbf{h}(l) \mid \mathbf{h}^{(k)}(l-1), \mathbf{y}(l)\big) &= p\big(\mathbf{h}(l) \mid \mathbf{h}^{(k)}(l-1), \mathbf{y}(l)\big) \\ &= \frac{p\big(\mathbf{y}(l) \mid \mathbf{h}(l), \mathbf{h}^{(k)}(l-1)\big)\, p\big(\mathbf{h}(l) \mid \mathbf{h}^{(k)}(l-1)\big)\, \cancel{p\big(\mathbf{h}^{(k)}(l-1)\big)}}{p\big(\mathbf{y}(l) \mid \mathbf{h}^{(k)}(l-1)\big)\, \cancel{p\big(\mathbf{h}^{(k)}(l-1)\big)}} \end{aligned} \tag{4.39}$$

Substituting into (4.38) yields a simple weight update equation.

$$w^{(k)}(l) \propto p\big(\mathbf{y}(l) \mid \mathbf{h}^{(k)}(l-1)\big) w(l-1)^{(k)}$$

The required distribution can be calculated using the Chapman-Kolmogorov equation (4.33). In many problems, the importance sampling distribution can be derived from the state and observation equations.

4.6.3 Degeneracy

Even with the optimal choice, as the particle filter evolves, degeneracy — small K_{eff} — will occur. A typical threshold for K_{eff} is $K/2$. The typical way of creating a better estimate of the conditional distribution is *resampling*, the technique that underlies many statistical techniques such as the bootstrap. The idea is very simple. Suppose you had an empirical estimate of a probability distribution; what would it mean to generate random variables according to it? The value of each random variable would equal one of the values used in the estimate. The probability that any given value would be selected equals its weight. Note that this procedure amounts to "sampling with replacement:" the probability of choosing a value does not change just because it has been chosen in previous simulation trials. Consequently, small probability values, which are the source of the degeneracy, are unlikely to be chosen but high-probability values will be repeated. The same number of resampled particles, K, as in the degenerate distribution are chosen. In extreme cases where resampling occurs many times, only a few values survive, leading to what is known as *sample impoverishment*.

4.6.4 Smoothing Estimates

So far, the particle filter algorithm can be used to produce filtering estimates and, with simple modification, prediction estimates. Smoothing estimates provide smaller error but can only be used in a batch processing mode in which you have all the observations before producing estimates.

To produce a particle filter for smoothing estimates, we use the weights produced at the final observation time to produce recursively an estimate of the conditional distribution $p\big(\mathbf{h}(l) \mid \mathbf{y}(0:l)\big)$ for previous times according to

$$p\big(\mathbf{h}^{(j)}(l), \mathbf{h}^{(k)}(l-1) \mid \mathbf{y}(0:l)\big) = p\big(\mathbf{h}^{(j)}(l) \mid \mathbf{y}(0:l)\big) \frac{p\big(\mathbf{h}^{(j)}(l) \mid \mathbf{h}^{(k)}(l-1)\big) w^{(k)}(l-1)}{\sum_k p\big(\mathbf{h}^{(j)}(l) \mid \mathbf{h}^{(k)}(l-1)\big) w^{(k)}(l-1)} \tag{4.40a}$$

$$p\big(\mathbf{h}^{(j)}(l-1) \mid \mathbf{y}(0:l)\big) = \sum_{k=1}^{K} p\big(\mathbf{h}^{(j)}(l), \mathbf{h}^{(k)}(l-1) \mid \mathbf{y}(0:l)\big) \tag{4.40b}$$

The idea is to initialize the recursion with $p\big(\mathbf{h}^{(j)}(l) \mid \mathbf{y}(0:l)\big) = w^{(j)}(l)$, the weights obtained at the time N of the last observation. The conditional distribution $p\big(\mathbf{h}^{(j)}(l) \mid \mathbf{h}^{(k)}(l-1)\big)$ is defined by the state-space equation (either (4.29) or (4.31)). The weights computed in the forward direction (after resampling) are reused. The result of the first equation (4.40a) is used in the second.

4.7 Spectral Estimation

One of the commonest requirements in statistical signal processing systems, both explicitly and implicitly, is estimating a signal's Fourier spectrum. Because the covariance function and the power spectrum are Fourier Transform pairs, estimation of one is related to estimation of the other. Thus, the following discussion frequently alternates between these two quantities.

The *resolution* of any spectral estimate is the degree to which it can express spectral detail. For example, if the spectrum is smooth (like that arising from first-order signals), little resolution is demanded of the spectral analysis. Conversely, signals consisting of sinusoids that give rise to spectra having abrupt changes require much more resolution. The more usual criteria of assessing the quality of an estimate (bias, consistency, etc..) are subsumed under spectral resolution considerations.

Resolution generally depends on two factors: the variance of the estimate and the degree of spectral smoothing implicit in the spectral estimation algorithm. A large variance may well mean that differences in spectral amplitude at two frequencies cannot be easily distinguished because of large statistical variations. Although the various spectral estimation algorithms do have very different statistical properties, they all share the property that variance depends *directly* on the spectrum being estimated. Thus, to characterize statistical errors, a general understanding of the spectrum's shape and how it varies with frequency, what we term *spectral structure*, is required to specify criteria on variability. As more data are incorporated into the spectral estimate, the variance should decrease (the estimate is consistent); for consistent estimators, the variance is ultimately limited by the amount of data available. *Spectral smoothing* expresses the detail that can be gleaned from an estimate when the statistical errors are small. An algorithm's asymptotic resolution is determined entirely by the range of frequencies that influence the estimate at each frequency. In many algorithms, this range is dependent on spectral structure and on the variance we demand: Algorithms having high asymptotic resolution usually have large variances and *vice versa*. A kind of "Catch-22" emerges: We have to know the spectrum to understand an algorithm's characteristics, but we don't know the spectrum; that's why we're trying to estimate it. Understanding this dilemma is the key to discerning effective spectral estimation techniques.

The various algorithms used in spectral estimation fall into two broad categories: *nonparametric* and *parametric*. Some idea of the spectral resolution necessary to reveal the spectrum's structure is required not only to select between algorithms in the two categories but also to select an algorithm within the category.

Nonparametric. We require no strong preconceived notions about the nature of a signal that can be successfully subjected to nonparametric spectral analysis. Relatively straightforward application of the discrete Fourier Transform characterizes these algorithms. They are by far the best understood; examples are the periodogram and Bartlett's procedure.

Parametric. Parametric algorithms assume we know a great deal *a priori* about a spectrum's structure. They are based on an explicit model for signal generation with the model's only uncertainty being the values of a few parameters. Thus, we must know the model's structure, and given this structure, the spectral estimation problem becomes one of parameter estimation. If the spectrum does fall into the assumed class, the estimate is very accurate (high resolution is achieved). If not, the estimate does not just have poor resolution, it can be misleading.

4.7.1 Periodogram

Let $s(n)$ denote a zero-mean, stationary, stochastic sequence having covariance function $K_s(m) = \mathsf{E}[s(n)s(n+m)]$ and power spectrum $\mathscr{S}_s(f)$, which are related to each other as

$$\mathscr{S}_s(f) = \sum_{m=-\infty}^{+\infty} K_s(m) e^{-j2\pi fm}$$

$$K_s(m) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \mathscr{S}_s(f) e^{+j2\pi fm}\, df$$

Because the Fourier Transform is linear, a linear estimate of one is equivalent to some linear estimate of the other. This relationship does *not* imply that they share statistical characteristics, however. This subtle point is beautifully exemplified by the periodogram.

Covariance-based spectral estimation. We observe the signal over the finite interval $[0, D-1]$. The covariance function can be estimated by approximating the averaging implicit in its definition with

$$\widehat{K}_s(m) = \frac{1}{D} \sum_{n=0}^{D-|m|-1} s(n)s(n+m), \quad 0 \le |m| \le D-1$$

The upper limit on this sum arises because of the data's finite extent: The signal $s(n+m)$ is provided for values of the index n ranging from $-m$ to $D-m-1$. Beyond lag $D-1$, this simple estimate provides no value. The expected value of this estimate is easily seen to be

$$\mathsf{E}[\widehat{K}_s(m)] = \left(1 - \frac{|m|}{D}\right) K_s(m)$$

Thus, this estimate is biased, but asymptotically unbiased. More important, bias differs for each lag, the smaller lags having much less bias than the larger ones. This effect is traced to the number of signal values that contribute to the estimate for a given lag: $D-m$ terms for lag m. Usually, we compute an average by summing a few terms, then dividing by the number of terms in the sum. The estimate of the covariance function given previously normalizes the sum by D no matter how many terms contribute to the sum; this choice leads to the bias. The variance of the estimate is given approximately by*

$$\mathsf{var}[\widehat{K}_s(m)] \approx \frac{1}{D} \sum_{n=-(D-m-1)}^{D-m-1} \left(1 - \frac{m+|n|}{D}\right) [K_s^2(n) + K_s(n+m)K_s(n-m)]$$

Consistent with the averaging interpretation just given, we find that the variance of the covariance function's estimate is bigger at large lags than at the small ones. Assuming the covariance function is square-summable, this summation remains finite as $D \to \infty$; hence the variance of the covariance estimate is proportional to $1/D$, and the estimate is consistent.

One way of estimating the power spectrum is to compute the Fourier Transform of the covariance function's estimate.

$$\widehat{\mathscr{S}_s}(f) = \sum_{m=-(D-1)}^{D-1} \widehat{K}_s(m) e^{-j2\pi fm}$$

This spectral estimate can be related directly to the observations by substituting the expression for the covariance estimate.

$$\widehat{\mathscr{S}_s}(f) = \frac{1}{D} \sum_{m=-(D-1)}^{D-1} \sum_{n=0}^{D-|m|-1} s(n)s(n+m)e^{-j2\pi fm}$$

The summation limits result from the finite extent of the observations. We can express the availability of data directly by replacing the signal by the *windowed signal* $w_D^R(n)s(n)$, where $w_D^R(n)$ is the rectangular or "boxcar" window of duration D.

$$w_D^R(n) = \begin{cases} 1 & 0 \le n \le D-1 \\ 0 & \text{otherwise} \end{cases}$$

The limits on the summations now become infinite with the result

$$\widehat{\mathscr{S}_s}(f) = \frac{1}{D} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} w_D^R(n) w_D^R(n+m) s(n)s(n+m)e^{-j2\pi fm}$$

*The reason for the approximation is the need for the *fourth* moment of the stochastic sequence s. This computation is dependent on the signal's joint amplitude distribution (see Problem 4.43). This approximation, derived under a Gaussian assumption, also applies to other distributions [52: 171–89].

A simple manipulation of the summand yields in an insightful re-expression.

$$\widehat{\mathscr{S}_s}(f) = \frac{1}{D} \sum_{n=-\infty}^{\infty} w_D^R(n) s(n) e^{+j2\pi fn} \sum_{l=-\infty}^{\infty} w_D^R(l) s(l) e^{-j2\pi fl}$$

The latter sum does not depend on n; thus, we find the spectral estimate to be the product of the windowed signal's Fourier Transform and its conjugate. Defining $S_D(f)$ to be the Fourier Transform of the windowed signal, this power spectrum estimate, termed the *periodogram*, is more succinctly written as [10];[71: 130–64];[77: 730–36]

$$\boxed{\begin{aligned} S_D(f) &= \sum_{n=-\infty}^{\infty} w_D^R(n) s(n) e^{-j2\pi fn} \\ \widehat{\mathscr{S}_s}(f) &= \frac{1}{D} |S_D(f)|^2 \end{aligned}}$$

The periodogram is, expectedly, biased because it is the Fourier Transform of a biased estimate of the covariance function. The expected value of the periodogram equals the Fourier Transform of the triangularly windowed covariance function.

$$\mathsf{E}[\widehat{\mathscr{S}_s}(f)] = \sum_{m=-(D-1)}^{D-1} \left(1 - \frac{|m|}{D}\right) K_s(m) e^{-j2\pi fm}$$

This window is colloquially termed a "rooftop" window; the literature terms it a *triangular* or *Bartlett* window.

$$w_{2D-1}^T(m) = 1 - \frac{|m|}{D}, \quad m = -D, \ldots, D$$

This lag-domain window is a consequence of the rectangular window we applied to the data. In the frequency domain, the periodogram's expected value is the convolution of the actual power density spectrum with the Bartlett window's Fourier transform.

$$\boxed{\mathsf{E}[\widehat{\mathscr{S}_s}(f)] = \int_{-\frac{1}{2}}^{\frac{1}{2}} \mathscr{S}_s(\alpha) W_{2D-1}^T(f-\alpha)\, d\alpha}$$

The Fourier Transform $W_{2D-1}^T(f)$ of the Bartlett window is known as the Féjer kernel.

$$W_D^T(f) = \frac{1}{D} \left(\frac{\sin \pi fD}{\sin \pi f}\right)^2$$

Window selection. These results derived for the rectangular window are easily generalized to other window shapes: Applying a window $w_D(n)$ to the data is equivalent to applying the window $w_{2D-1}^c(m)$ to the covariance function, where the covariance-domain window equals the autocorrelation function of the data window. The relationship of the data window $w_D(n)$ to the periodogram's expected value $\mathsf{E}[\widehat{\mathscr{S}_s}(f)]$ is summarized as

$$\boxed{\begin{aligned} w_{2D-1}^c(m) &= \frac{1}{D} \sum_{n=-\infty}^{\infty} w_D(n) w_D(n+m) \\ W_{2D-1}^c(f) &= \sum_{m=-\infty}^{\infty} w_{2D-1}^c(m) e^{-j2\pi fm} \\ \mathsf{E}[\widehat{\mathscr{S}_s}(f)] &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \mathscr{S}_s(\alpha) W_{2D-1}^c(f-\alpha)\, d\alpha \end{aligned}} \tag{4.41}$$

Because the actual power spectrum is convolved with the kernel $W_{2D-1}^c(\cdot)$, the periodogram is a *smoothed* estimate of the power spectrum. Only those aspects of the spectrum that vary over a range of frequencies *wider* than the width of the kernel are noticeable. This width can be defined much like a mainlobe as the distance between the smallest zeros surrounding the origin. For the Féjer kernel, this width is $4\pi/D$. The periodogram's bias thus has a very complicated structure, being neither additive nor multiplicative.

4.7.2 Short-Time Fourier Analysis

The key theoretical framework that describes computing spectra over finite time intervals is the *short-time Fourier Transform*. Calculating this quantity means that we apply to a signal at time t a window of duration D, then evaluate the Fourier Transform of the product.

$$X_m(t,f) \equiv \int_t^{t+D} w(\tau - t) X_m(\tau) e^{-j2\pi f\tau}\, d\tau \tag{4.42}$$

Here, $w(t)$ denotes the finite-duration window defined over $[0,D]$. The window's duration defines the frequency resolution of the short-time Fourier analysis: Because $X_m(t,f)$ equals the Fourier Transform of the product between the sensor signal and the window origined at t, the signal's spectrum is smoothed by the window's spectrum. For the short-time spectrum calculated at one time instant to approximate the signal's spectrum well, we should choose the window's duration so that the spectral smoothing function's width is narrow compared with the signal's spectral variations. Reasonable windows have Fourier Transforms that resemble a single narrow spike; examples are shown in Fig. 4.10 {118}. Simple manipulations yield an equivalent representation of the short-time Fourier Transform.

$$X_m(t,f)e^{j2\pi ft} = \int_0^D w(\tau) X_m(t+\tau) \exp\{-j2\pi f\tau\}\, d\tau$$

This representation means that $X_m(t,f)$ is a complex-valued lowpass signal that approximates the "local" spectrum of the sensor's output at time t and frequency f.

Similar to the definition given in Eq. 4.42 for the short-time Fourier Transform of a continuous-time signal, the short-time Fourier Transform of a sampled sensor signal $X_m(n)$ is given by

$$X_m(n_0,f) = \sum_{l=n_0}^{n_0+D-1} w(l-n_0) X_m(l) \exp\{-jl2\pi fT\}, \quad f \text{ in Hz} \tag{4.43}$$

The D values of X_m following the time-sample n_0 are used to calculate $X_m(n_0,f)$. The function $w(n)$ is a time-domain window equaling zero outside the interval $[0,D-1]$; thus, D represents the window's duration. Note that the product fT denotes the usual digital frequency variable that has units of radians per sample period T. A simple change in the summation variable shows that the short-time Fourier Transform has the related expression

$$X_m(n_0,f)e^{jn_0 2\pi fT} = \sum_{l=0}^{D-1} w(l) X_m(n_0+l) \exp\{-jl2\pi fT\} \tag{4.44}$$

The shape of the window function $w(n)$ influences the resolution and detailed structure of the spectrum $X_m(n_0,f)$. Its Fourier Transform $W(f)$ effectively smoothes the desired but unobtainable spectrum $X_m(f)$ to give us the short-time spectrum.

The relation of the periodogram's expected value to the convolution of the actual spectrum with the window's magnitude-squared Fourier Transform (Eq. 4.41) expresses the smoothing characteristics of periodogram-based spectral estimates. Using language gleaned from conventional beamforming, the Fourier Transform of a rectangular window $\sin(\pi fD)/\sin(\pi f)$ has a very narrow mainlobe and large ripples (sidelobes) elsewhere. The sidelobes are directly related to the abrupt beginning and termination of the window. We can use window shapes other than the rectangular to improve these smoothing characteristics. Windows having reduced sidelobes have *tapered* ends: They slowly decrease to 0 near the ends. As expected, this reduction is compensated by an increase in mainlobe width. For example, a Bartlett window may be applied to a frame: The mainlobe of its Fourier Transform has twice the width of the rectangular window's transform, but the sidelobes' amplitudes are greatly reduced. Window taper also suppresses the data that are unfortunate enough to occur at a frame's edges; they tend to be ignored in the spectral estimate. Some well-known window

Window	*Formula*[a]	*FWHM*[b]	*Highest Sidelobe Level* (dB)
Rectangular	1	$7.60/D$	-13
Triangular	$1-\frac{\lvert n-D_{1/2}\rvert}{D_{1/2}}$	$11.2/D$	-27
Hamming	$\left(0.54-0.46\cos\frac{2\pi n}{D}\right)$	$11.4/D$	-43
Dolph-Chebyshev[c]	$W(k)\propto\begin{cases}\cos\{D\cos^{-1}[\theta(k)]\}, & \lvert\theta(k)\rvert\le 1\\ \cosh\{D\cosh^{-1}[\theta(k)]\}, & \lvert\theta(k)\rvert>1\end{cases}$ where $\theta(k)=\beta\cos\pi k/D$	$11.6/D$	-50
Kaiser[d]	$I_0\left[\pi\alpha\sqrt{1-\left(\frac{n-D_{1/2}}{D_{1/2}}\right)^2}\right]/I_0(\pi\alpha)$	$12.5/D$	-46
Hanning	$\frac{1}{2}\left(1-\cos\frac{2\pi n}{D}\right)$	$12.6/D$	-32

[a]The formula for each window presumes that they are defined over the interval $0\le n<D$ with $D_{1/2}=(D-1)/2$.

[b]Full-width half-maximum measure of mainlobe width.

[c]$W(k), k=0,\ldots,D-1$ is the discrete Fourier Transform of the window function. For the purposes of this formula, $\cosh^{-1}(x)$ is given by $\ln\left(|x|+\sqrt{x^2-1}\right)$. The parameter β equals $\cosh[(\cosh^{-1}10^{\alpha})/D]$, where α is the number of decades the sidelobe heights lie below the mainlobe height. In this table, the numerical values were obtained with $\alpha=2.5$.

[d]The parameter α controls the mainlobe width and sidelobe levels. The larger it is, the wider the mainlobe but smaller the sidelobes. The numbers presented here result from $\alpha=2.0$.

Table 4.1: Characteristics of Commonly Used Windows [adapted from Harris].

functions along with their spectral smoothing functions (their Fourier Transforms) are shown in Fig. 4.10. A tabular display of their characteristics is provided in Table 4.1.

In a general sense, the effect of windowing on spectral resolution is *independent* of the shape of the window. As more and more data are used, the ultimate resolution of periodogram-based spectral estimates is directly dependent on frame duration, with window choice changing the constant of proportionality via mainlobe width. All windows produce sidelobes of varying structure; sidelobe structure affects *spectral leakage*: the influence on the estimate at a particular frequency by remote portions of the spectrum.

Statistical characteristics. Despite the expected value of the periodogram being a seemingly reasonable approximation to a signal's power spectrum and despite the fact that the estimate of the covariance function is asymptotically unbiased and consistent, *the periodogram does not converge to the power spectrum* [52: 222–23];[63: 260–65]: The variance of the power spectrum does *not* tend to zero as the number of observations increases. One simple example of this result is the white noise case. Let's calculate the covariance between the power spectrum estimates at the frequencies f_1, f_2. First of all,

$$\mathsf{E}[\widehat{\mathscr{S}_s}(f_1)\widehat{\mathscr{S}_s}(f_2)]=\left(\frac{1}{D}\right)^2\sum_{k,l,m,n=0}^{D-1}\mathsf{E}[s(k)s(l)s(m)s(n)]e^{j2\pi f_1(l-k)}e^{j2\pi f_2(n-m)}]$$

Because of the Gaussian assumption, we can use the fourth-moment property of jointly Gaussian random variables {10} to find that

$$\mathsf{E}[s(k)s(l)s(m)s(n)]=\sigma_s^4[\delta(l-k)\delta(n-m)+\delta(l-m)\delta(k-n)+\delta(l-n)\delta(k-m)]$$

Consequently,

$$\mathsf{E}[\widehat{\mathscr{S}_s}(f_1)\widehat{\mathscr{S}_s}(f_2)]=\sigma_s^4\left[1+\left(\frac{\sin\pi(f_1+f_2)D}{D\sin\pi(f_1+f_2)}\right)^2+\left(\frac{\sin\pi(f_1-f_2)D}{D\sin\pi(f_1-f_2)}\right)^2\right]$$

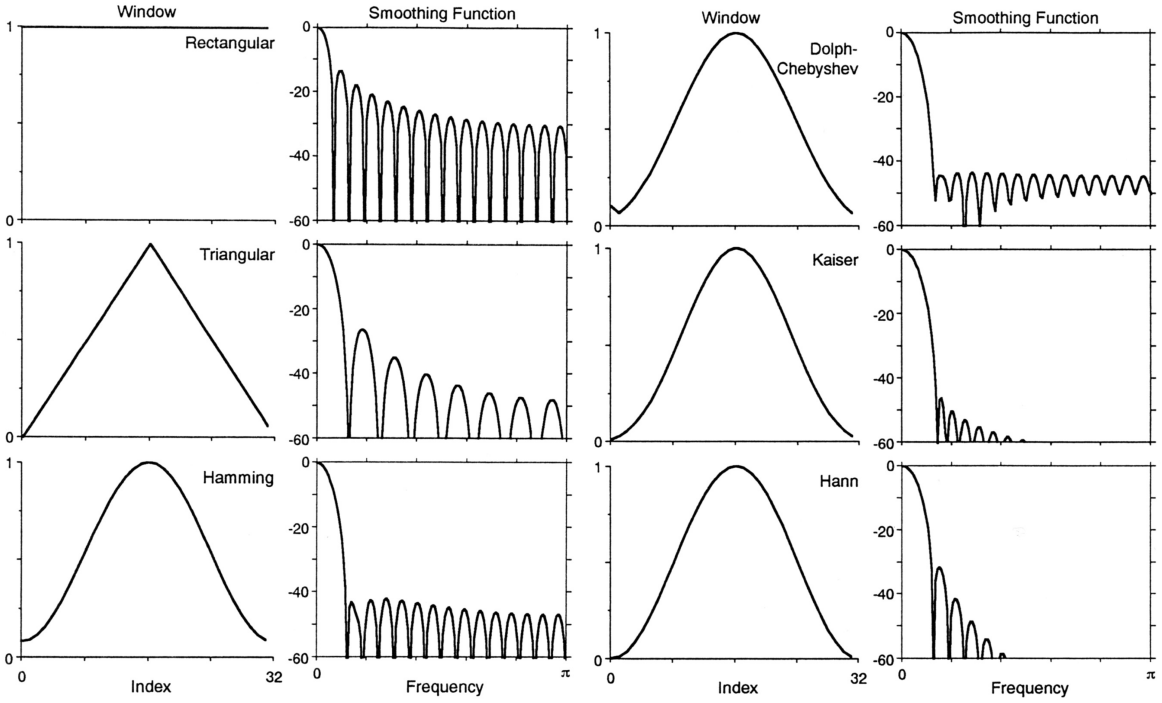


Figure 4.10: Window functions frequently used in spectral estimation are shown in the left column and their Fourier Transforms—smoothing functions—in the right. The spectral amplitudes are plotted in dB *re* the spectral amplitude at the origin. The window duration was $D = 32$. The latter quantities, when convolved with the true spectrum, yield the expected value of the Bartlett spectral estimate. For those frequency ranges where no spectral values are shown, the smoothing functions are less than -60 dB. The parameters of the Dolph-Chebyshev and Kaiser windows are those used in Table 4.1. The ideal of a narrow mainlobe and small sidelobes is never achieved as these examples show; a compromise is always required in practice.

which means the covariance equals

$$\mathsf{cov}[\widehat{\mathscr{S}_s}(f_1)\widehat{\mathscr{S}_s}(f_2)] = \sigma_s^4 \left[\left(\frac{\sin \pi(f_1+f_2)D}{D \sin \pi(f_1+f_2)} \right)^2 + \left(\frac{\sin \pi(f_1-f_2)D}{D \sin \pi(f_1-f_2)} \right)^2 \right]$$

First of all, this result means that the spectral estimates at different frequencies are correlated, with the correlation decreasing as $1/D^2$. To find the variance, we let $f_1 = f_2$ with the result that

$$\mathsf{var}[\widehat{\mathscr{S}_s}(f)] = \sigma_s^4 \left[1 + \left(\frac{\sin 2\pi f D}{D \sin 2\pi f} \right)^2 \right]$$

As $D \to \infty$, the second term goes to zero whenever $f \neq \{0, \frac{1}{2}\}$,* leaving $\lim_{D\to\infty} \mathsf{var}[\widehat{\mathscr{S}_s}(f)] = \sigma_s^4$, the square of the signal's power spectrum. More generally, the periodogram's asymptotic variance, for Gaussian signals, tends at any frequency toward the square of the power density spectrum (see Problem 4.43).

$$\lim_{D\to\infty} \mathsf{var}[\widehat{\mathscr{S}_s}(f)] \propto \mathscr{S}_s^2(f)$$

Because the asymptotic variance is not 0, the periodogram is *not* a consistent estimate of the power spectrum. This result applies to non-Gaussian signals as well. In estimation theory, the periodogram is perhaps the most famous example of an inconsistent, yet asymptotically unbiased estimator. This lack of convergence

*This term equals one at these frequencies.

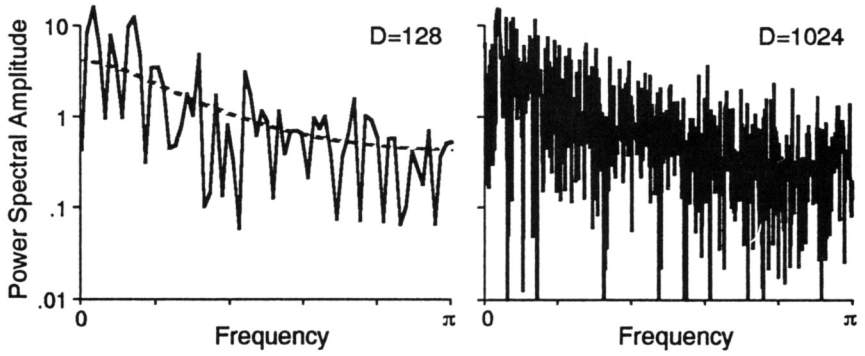


Figure 4.11: Periodograms were computed from the example signal used in the Wiener and adaptive filtering sections of this chapter. A rectangular window was applied in each case with the spectrum in the left panel computed from 128 data points and from 1,024 points for the one on the right. Note how the estimate's variability does not stabilize as more data are incorporated.

is illustrated in Fig. 4.11.

Why shouldn't the periodogram converge? The reason lies in the expression of the power spectrum as an integral transformation of the estimated covariance function. Recall that the variances of the covariance function's values at higher lags were large. These terms are added in the Fourier sum, and their total variability is never compensated by the normalization factor. Realizing this flaw in the periodogram is the key to finding alternate spectral estimates that do converge.

Bartlett's procedure. To obtain a consistent spectral estimate from the empirical covariance function, either higher lags must not be used in the computations or more data should contribute toward their average. With the latter idea in mind, from the total of L observations, frame the data into K sections of equal duration D ($L = KD$). The *Bartlett procedure* forms a power spectrum estimate by averaging the spectra computed from each frame [7];[71: 153–58];[94].*

$$\widehat{\mathscr{S}_s}(f) = \frac{1}{K}\sum_{k=0}^{K-1}\widehat{\mathscr{S}_s^{(k)}}(f)$$

$$\widehat{\mathscr{S}_s^{(k)}}(f) = \frac{1}{D}\left|S_D^{(k)}(f)\right|^2$$

$$S_D^{(k)}(f) = \sum_{n=0}^{D-1} w_D(n)s(kD+n)e^{-j2\pi fn}$$

In this case, periodograms are computed from each frame; conceptually, any spectral estimate could be averaged in this way. Assuming that the frames are mutually independent, the variance of the Bartlett estimate equals the sum of the individual variances divided by K.

$$\mathsf{var}[\widehat{\mathscr{S}_s}(f)] \propto \frac{1}{K}\mathscr{S}_s^2(f) \tag{4.45}$$

Thus, as $L \to \infty$, maintaining fixed duration frames implies that $K \to \infty$ so that the Bartlett spectral estimate is consistent. The non-convergent periodogram example is reexamined in Fig. 4.12 with the Bartlett procedure.

When windows other than the rectangular are used, data aligned with the ends of our somewhat arbitrarily selected frames do not contribute to the spectral estimate. To counteract this arbitrary stress on portions of the data, frames are overlapped (Fig. 4.13): The separation between successive windows, the stride, is less

*The quantity $S_D^{(k)}(f)$ corresponds exactly to the short-time Fourier Transform $S(kD,f)e^{jkD2\pi f}$ defined previously (Eq. 4.44 {116}). The more streamlined current notation is employed to simplify the equations for Bartlett's procedure. Note that the linear phase shift $e^{jkD2\pi f}$ owing to the signal's nonzero time origin is removed by the magnitude operation of the second equation.

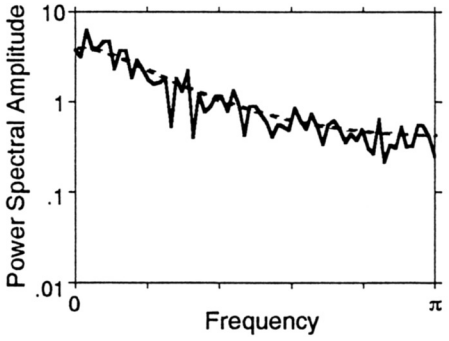


Figure 4.12: The data used in the right panel of the periodogram example (Fig. 4.11) are now used in Bartlett's procedure. The duration of each frame is 128 points, and the Fourier Transform of each frame has the same length. Eight periodograms are averaged to produce this spectrum; thus, a total of 1,024 data points were used in this computation and in the last portion of the periodogram example. The marked decrease in statistical variability is evident. Also note that the variability has roughly the same amplitude at each frequency. This effect is because of the relation of the spectral estimate's variance to the square of the actual spectral amplitude and to the logarithmic vertical scale.

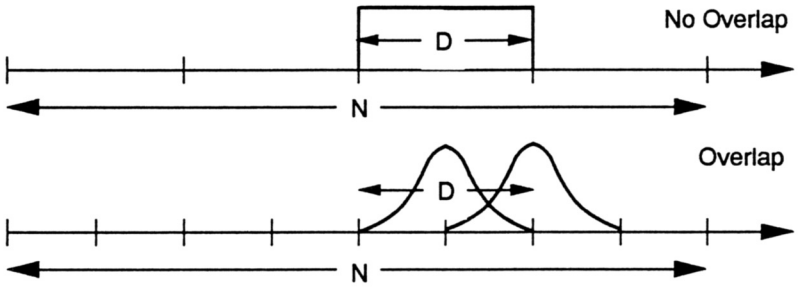


Figure 4.13: In Bartlett's spectral estimation procedure, the observations are subdivided into frames of length D. The periodogram is computed over each frame, and these spectra are averaged to produce a spectral estimate. The frames do not usually overlap, as shown in the upper portion, where a rectangular window is used. A 50% overlap or, said another way, 50% stride, is typical when one of the gently tapered windows (such as the Hanning window shown here) is used.

---

than the window's duration. A 50% overlap is commonly used. With overlap, many more frames are obtained from a given set of data. This increased number does *not* mean that the variance of the estimate is reduced proportionally. The variance of the spectra resulting from Bartlett's procedure is inversely proportional to the number of *statistically independent* frames; overlapping frames introduces a dependence between them. The number of statistically independent frames equals the number of nonoverlapping frames. The effect of window shape and overlapping frames is illustrated in Fig. 4.14.

Consistency in this spectral estimate has been obtained at the expense of increased bias. The duration of a frame is much smaller than the total, and this duration increases the effective width of the kernel that smooths the spectrum. The variance is proportional to $1/K$, and the width of the smoothing kernel is proportional to $1/D$. The product of these is thus proportional to $1/L$, which is inversely related to the amount of available data. In this way, a fundamental tradeoff in spectral estimation emerges: *The product of the degree of smoothing of the spectral estimate and of the estimate's variance is a constant*. In statistical terms, the Bartlett estimate may be consistent, but it is *not* (even) asymptotically unbiased because of spectral smoothing. Such is the nature of the fundamental tradeoff of spectral estimation in the Bartlett procedure.

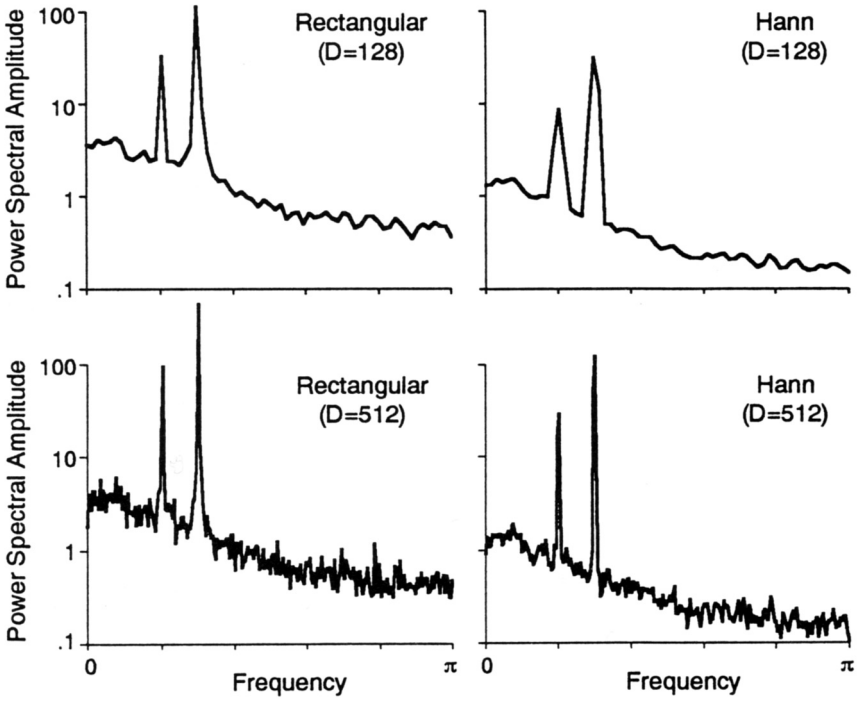


Figure 4.14: The observations consist of two sinusoids having frequencies $2\pi \times 0.1$ and $2\pi \times 0.15$ and additive noise having first-order lowpass spectral characteristics. The power of the higher frequency sinusoid is four times that of the lower. This additive noise is the same as that presented in the adaptive filtering examples. The number of observations available equals 8,192. The left column portrays the Bartlett spectral estimate that uses a rectangular data window with no overlap, the right column a Hanning window with a 50% overlap. The top row uses 128-point component periodograms, the bottom 512. Because of the smaller energy of the Hanning window, the spectra produced by the Hanning window are scaled relative to the ones produced by the rectangular window (both windows had unity height).

4.7.3 Minimum Variance Spectral Estimation

Bartlett's procedure yields a consistent spectral estimate by averaging periodograms computed over short frames of the observations. An alternate approach is to compute the covariance function over a few lags, but compute it over the entire data set. After applying any tapered window to this estimate, computing the Fourier Transform results in a spectral estimate that does not differ greatly from that yielded by Bartlett's procedure. A more interesting use of the covariance estimate is the so-called *minimum variance* spectral estimate [17];[18];[71: 350–57], which uses the covariance function in the solution of a somewhat artificial optimization problem. The resulting estimate is a *nonlinear* function of the data; hence its characteristics are not readily analyzed. Suffice it to say that its spectral resolution is greatly superior to Bartlett's procedure, but much more computation is needed to produce the estimate.

Similar to the *FIR* Wiener filter problem, assume that the observations are sent through a finite-duration filter that passes a known signal with unity gain while minimizing the contribution to the filter's output of all other components in the observations. Let the known signal be denoted by the column matrix $\mathbf{s}$ and the filter's unit-sample response by $\mathbf{h}$. Require the filter's output to the signal component $\mathbf{h}'\mathbf{s}$ to be unity while minimizing the mean-squared value of the entire output $\mathsf{E}[|\mathbf{h}'\mathbf{X}|^2] = \mathbf{h}'\mathbf{K}_X\mathbf{h}$. The minimum variance approach imposes the constrained optimization problem [53, 64]

$$\min_{\mathbf{h}} \mathbf{h}'\mathbf{K}_X\mathbf{h} \quad \text{subject to } \mathbf{h}'\mathbf{s} = 1$$

This kind of constrained optimization problem is easily solved using Lagrange multipliers (see §3.2.1 {51}).

The filter that yields the minimum variance* signal estimate has unit-sample response

$$\boxed{\mathbf{h}_\diamond = \frac{\mathbf{K}_X^{-1}\mathbf{s}}{\mathbf{s}'\mathbf{K}_X^{-1}\mathbf{s}}}$$

To apply this somewhat arbitrary problem and its solution to spectral estimation, first note that the solution depends on *only* the observations' covariance matrix over the few lags that correspond to the assumed filter duration. Given an estimate of this covariance matrix, we can approximate the optimal filter by using the estimate in the formula given earlier. To focus this reasoning further, let the "signal" be a complex exponential corresponding to the frequency at which we want to estimate the spectrum: $\mathbf{s} = \mathbf{e}(f) = \text{col}\left[1, e^{j2\pi f}, e^{j2\cdot 2\pi f}, \ldots, e^{j(M-1)2\pi f}\right]$. Thus, the length-$M$ filter passes this signal with unity gain—performs a spectral analysis—while minimizing the mean-squared value of the filter's output (asking for the maximum resolution possible by rejecting the influence of all other spectral components). In this way, a somewhat artificial optimal filtering problem has been recast as a potentially interesting spectral estimation procedure. The power spectrum would correspond to the filter's output power; this quantity equals the mean-squared value of the filtered observations. Substituting the optimal unit-sample response into the expression for the output's mean-squared value results in the minimum variance spectrum.

$$\mathscr{S}_X^{\text{MV}}(f) = \mathsf{E}\left[\left|\mathbf{h}_\diamond' \mathbf{X}\right|^2\right] = \left[\mathbf{e}'(f)\mathbf{K}_X^{-1}\mathbf{e}(f)\right]^{-1} \tag{4.46}$$

The minimum variance spectral estimate is obtained by replacing the covariance matrix by its estimate.

$$\boxed{\begin{aligned} \widehat{\mathscr{S}}_X^{\text{MV}}(f) &= \left[\mathbf{e}'(f)\widehat{\mathbf{K}}_X^{-1}\mathbf{e}(f)\right]^{-1} \\ \widehat{K}_{i,j}^X &= \frac{1}{D}\sum_{n=0}^{D-|i-j|-1} X(n)X(n+|i-j|), \quad 0 \le i,j < M \end{aligned}} \tag{4.47}$$

This spectral estimate is not a simple Fourier Transform of the covariance estimate: The covariance matrix estimate is formed from length-D observations, its inverse computed and used in a quadratic form that is approximately equivalent to the Fourier Transform of a matrix, and the reciprocal evaluated. Conceptually, the quadratic form is the kernel's Fourier Transform. The Fourier Transform of a column matrix is concisely expressed as $\mathbf{e}'(f)\mathbf{X}$; the combination of pre-multiplying and post-multiplying a matrix by $\mathbf{e}$ amounts to a Fourier Transform (see Problem 4.47). Fast Fourier Transform algorithms can be used to transform a column matrix, but not a square one unless the matrix has many symmetries. Although a covariance matrix may be Toeplitz as well as Hermitian, its inverse is only Hermitian, thereby obviating fast algorithms. Note that the number M of covariance function lags used in the estimate is usually much smaller than the number of observations D. In contrast to Bartlett's procedure, where many covariance estimates computed over short frames are averaged to produce a covariance estimate, here the entire set of observations is used to produce estimates for a similar number of lags. A Bartlett-procedure-like variation of estimating the covariance matrix is derived in Problem 4.46. Another consequence of $M \ll D$ is the number of frequencies at which the quadratic form can be evaluated to reflect the algorithm's resolution accurately. In most cases, the width of the smoothing function is comparable with that of the Féjer kernel implicit in rectangular data windows of length D. Despite the comparatively small size of the matrix, the quadratic form needs evaluation at *many* frequencies to capture the minimum variance algorithm's resolution. A computational note: The inverse of this matrix need not be computed directly. The minimum variance estimate is intimately related to the so-called autoregressive (AR) spectral estimate described in the next section (Eq. 4.50 {125}). This relationship not only expresses the deep relationship between two very different spectral estimators but also provides a computationally expedient method of calculating the minimum variance estimate.

*The observations are assumed to be zero mean. Thus, the mean-squared value and the variance are equal. Historically, even in nonzero mean cases the solution to the "minimum variance" problem is taken to be this minimum mean-squared value solution. So much for the name of an algorithm suggesting what aspects of the data are considered important or how the result was derived.

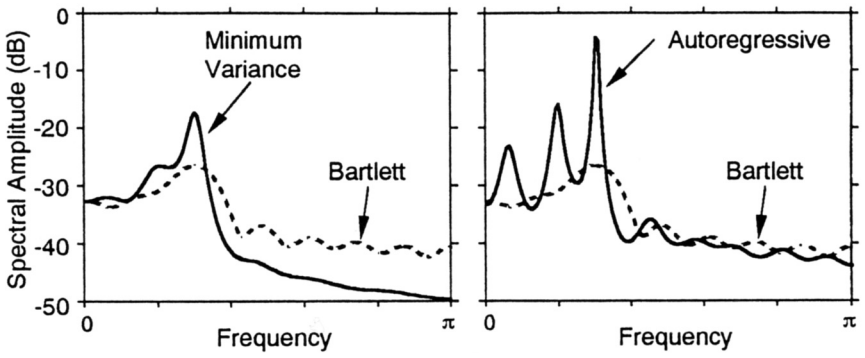


Figure 4.15: The plots depict minimum variance and *AR* spectral estimates computed from the same data ($D = 8,192$) used in Fig. 4.14 {121}. Only the *AR* spectrum has a peak at the frequency of the smaller sinusoid ($f = 0.1$); however, that peak is mitigated by the presence of a spurious peak located at a lower frequency. The size M of the covariance matrix is 16.

The minimum variance spectrum is related to the signal's power spectrum similarly to the periodogram: The minimum variance spectrum is a smoothed version of the true spectrum. To demonstrate this property, any spectral estimate $\widehat{\mathscr{S}_X}(f)$ given by the mean-squared value of weighted observations is related to the covariance function by

$$\mathsf{E}\left[\widehat{\mathscr{S}_X}(f)\right] = \mathsf{E}\left[\left|\mathbf{h}'\mathbf{X}\right|^2\right] = \sum_{k,l} h(k)K_X(k-l)h(l)$$

Using Parseval's Theorem, this expression for the spectrum at frequency f_0 is given by

$$\mathsf{E}\left[\widehat{\mathscr{S}_X}(f_0)\right] = \int_{-\frac{1}{2}}^{\frac{1}{2}} |H(f_0,f)|^2\, \mathscr{S}_X(f)\, df$$

$H(f_0,f)$ is the Fourier Transform of the filter's unit-sample response $\mathbf{h}$; for the minimum variance spectral estimate, this smoothing function is given by

$$H^{\mathrm{MV}}(f_0,f) = \frac{\mathbf{e}'(f)\mathbf{K}_X^{-1}\mathbf{e}(f_0)}{\mathbf{e}'(f_0)\mathbf{K}_X^{-1}\mathbf{e}(f_0)}$$

Because the minimum variance smoothing function is itself dependent on the data's characteristics, it changes form according to the shape of the signal's spectrum near the frequency of interest. In this way, the smoothing function depends directly on the value of the analysis frequency (f_0), not on the difference between f_0 and f. The smoothing kernel is dependent on the structure of the spectrum; *its resolution is therefore data dependent*.

When the observations are Gaussian, the variance of the minimum variance spectral estimate is given by [19]

$$\mathsf{var}[\widehat{\mathscr{S}}_X^{\mathrm{MV}}(f)] = \frac{1}{D-M+1}\,\mathsf{E}[\widehat{\mathscr{S}}_X^{\mathrm{MV}}(f)]$$

Thus, for fixed-size covariance matrices, the minimum variance spectral estimate is consistent. The estimate's expected value is proportional to the true minimum variance spectrum.

$$\mathsf{E}[\widehat{\mathscr{S}}_X^{\mathrm{MV}}(f)] = \frac{D-M+1}{D} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|H^{\mathrm{MV}}(f_0,f)\right|^2 \mathscr{S}_X(f)\, df$$

The minimum variance spectral estimate is therefore asymptotically unbiased and consistent. For a slightly larger variance than Bartlett's procedure, the minimum variance estimator provides much better resolution. Fig. 4.15 demonstrates this well.

4.7.4 Spectral Estimates Based on Linear Models

The most prevalent model for the generation of a signal is passing white noise through a linear system. The foundation of Wiener and Kalman filters, the same model can be used fruitfully in spectral estimation. By estimating model parameters, we can find the transfer function and thereby deduce the output's—the signal's—spectrum. Letting S denote the observations and W the input noise, the input-output relationship of a rational linear model is

$$S(n) = a_1 S(n-1) + \cdots + a_p S(n-p) + W(n) + b_1 W(n-1) + \cdots + b_q W(n-q) \tag{4.48}$$

In statistics, this relationship governs "autoregressive-moving average" (*ARMA*) models [71: 172–86]. The word "autoregressive" refers to models where the signal depends on itself in a linear fashion. "Moving average" refers to a linear combination of white-noise values. Thus, the model's autoregressive portion refers to the recursive aspect related to the poles and the moving average to the zeros of some linear system's transfer function. Spectral estimation based on this linear model for the signal amounts to estimation of the model's parameters. Given these values, the $ARMA(p,q)$ spectral estimate has the form

$$\widehat{\mathscr{S}}_S^{\text{ARMA}}(f) = \hat{\sigma}_W^2 \frac{\left|1 + \hat{b}_1 e^{-j2\pi f} + \cdots + \hat{b}_q e^{-j2\pi fq}\right|^2}{\left|1 - \hat{a}_1 e^{-j2\pi f} - \cdots - \hat{a}_p e^{-j2\pi fp}\right|^2}$$

AR spectral estimation. The *AR* method is by far the most popular *parametric* or *model-based* spectral estimators [71: 189–274]. Translating into signal processing terms, an autoregressive signal is produced at the output of an all-pole linear filter driven by white noise.

$$S(n) = a_1 S(n-1) + \cdots + a_p S(n-p) + W(n)$$

The parameter p denotes the model's *order*. One method of estimating the parameters of the *AR* spectral estimate was given in a previous section (Eq. 4.18 {92}). The *linear predictive* solution is found by finding the minimum mean-squared error one-step predictor.*

$$\frac{1}{\hat{\sigma}_W^2}\begin{bmatrix} 1 \\ -\hat{a}_1 \\ \vdots \\ -\hat{a}_p \end{bmatrix} = \left(\widehat{\mathbf{K}}_S\right)^{-1}\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The variance estimate $\hat{\sigma}_W^2$ is equal to the reciprocal of the first component of $\widehat{\mathbf{K}}_S^{-1}\boldsymbol{\delta}$, where $\boldsymbol{\delta} = \text{col}[1,0,\ldots,0]$. Because of the way the input term $W(n)$ enters the difference equation for the signal, the mean-squared prediction error equals $\hat{\sigma}_W^2$. The *AR* spectral estimate has the form

$$\widehat{\mathscr{S}}_S^{\text{AR}}(f) = \frac{\boldsymbol{\delta}^t \widehat{\mathbf{K}}_S^{-1}\boldsymbol{\delta}}{\left|\boldsymbol{\delta}^t \widehat{\mathbf{K}}_S^{-1}\mathbf{e}(f)\right|^2} = \frac{\hat{\sigma}_W^2}{\left|1 - \hat{a}_1 e^{-j2\pi f} - \cdots - \hat{a}_p e^{-j2\pi fp}\right|^2} \tag{4.49}$$

An example of this estimate is shown in Fig. 4.15 {123}. Another popular technique for estimating these parameters is Burg's method [71: 213–15].

The *AR* estimator is also known as the *maximum entropy* spectral estimate [12, 51]. How this name arose is an interesting story that recalls the relation between the power spectrum and the covariance function. The asymptotic resolution of the periodogram-based spectral estimates is directly related to how many lags of the covariance function can be estimated accurately: More lags imply higher resolution. Suppose $p+1$ lags can be well estimated. Can the higher lags be accurately *extrapolated* from the well-known ones in such a way

*Note the care used in denoting the inverse of the covariance matrix estimate rather than the estimate of the inverse. Directly estimating the inverse is touchy at best. We have discussed several methods of estimating the covariance matrix ({97} and {122}). The linear predictive literature has developed even more [69].

that the spectral resolution is increased? Seemingly, this approach is trying to "get something for nothing." Not any extrapolation, however, can result in a covariance function; the Fourier Transform of any potential covariance function must be real and nonnegative. Imposing this restriction greatly reduces the number of possible extrapolations, but an infinite number remain. From statistical arguments, one can show that the spectral estimate that coincides with the known values of the covariance function and corresponds to the most probable extrapolation of the covariance function maximizes the *entropy* $\mathscr{H}[\mathscr{S}]$ of the power spectrum $\mathscr{S}$.

$$\mathscr{H}[\mathscr{S}] = \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln \mathscr{S}(f)\, df$$

This framework defines a constrained maximization problem, the solution to which is a *AR* spectral estimate of the form given earlier! Thus, the *AR*, model-based technique holds the promise of greater resolution than Bartlett's procedure.

By noting that the squared magnitude of a sequence's Fourier Transform equals the Fourier Transform of its autocorrelation, we find an alternative to the denominator expression in the *AR* spectral estimate.

$$\left|1 - \hat{a}_1 e^{-j2\pi f} - \cdots - \hat{a}_p e^{-j2\pi f p}\right|^2 = \sum_{m=-p}^{p} R_{\hat{a}}^{\mathrm{AR}}(m) e^{-j2\pi f m}$$

$$R_{\hat{a}}^{\mathrm{AR}}(|m|) = \sum_{k=0}^{p-m} \hat{a}_k \hat{a}_{k+m}, m \geq 0$$

For the purposes of this re-expression, $\hat{a}_0 = -1$. This re-expression is not particularly useful as it stands: It provides no computational advantages. Its value lies in its relation to a similar expression for the minimum variance spectral estimate [75]. Because the solution for the minimum variance estimator (Eq. 4.47 {122}) also contains the inverse of the covariance matrix, the minimum variance spectral estimate can be related to the *AR* parameter estimates.

$$\boxed{\begin{aligned} \widehat{\mathscr{S}}_S^{\mathrm{MV}}(f) &= \frac{\hat{\sigma}_W^2}{\sum_{m=-p}^{p} R_{\hat{a}}^{\mathrm{MV}}(m) e^{-j2\pi f m}} \\ R_{\hat{a}}^{\mathrm{MV}}(|m|) &= \sum_{k=0}^{p-m} (p+1-m-2k) \hat{a}_k \hat{a}_{k+m}, m \geq 0 \end{aligned}} \tag{4.50}$$

The correlation function resembles an expression for a windowed version of the *AR* parameter estimates' autocorrelation. It does point out one of the ways that minimum variance and *AR* spectral estimates can be related. More important, because efficient methods exist for computing the *AR* parameter estimates, those algorithms can be used to compute the minimum variance estimate as well.

A critical issue in *AR* spectral estimation is value of the model order p. As model order is increased, more parameters become available to model the signal's spectrum, suggesting that mean-squared prediction error decreases. Using larger values of p, however, implies accurate knowledge of the covariance function at larger lags, which we know are less reliable statistically. Thus, a parsimonious criterion for selecting a value for p is required. Two techniques are most well-known: the *A* information criterion (*AIC*) [2] and the minimum description length (*MDL*) criterion [81, 83]. Each of these criteria seeks the model order that minimizes the sum of two terms: The first is the negative logarithm of the likelihood function, where the *AR* parameters are evaluated at their maximum likelihood estimates, and the second is a constant times the number of parameters $p+1$. The criterion functions differ only in the constant used in the second term [69];[71: 230–31].*

$$\boxed{\begin{aligned} AIC(p) &= -\ln \Lambda(\hat{\sigma}_W^2, \hat{a}_1, \ldots, \hat{a}_p) + p \\ MDL(p) &= -\ln \Lambda(\hat{\sigma}_W^2, \hat{a}_1, \ldots, \hat{a}_p) + \frac{1}{2} p \ln D \end{aligned}}$$

*Because we are only interested in the behavior of these criterion functions with variations in p, additive terms not depending on p are dropped as they do not affect the location of the minimum. These criterion functions have been so manipulated to their simplest form.

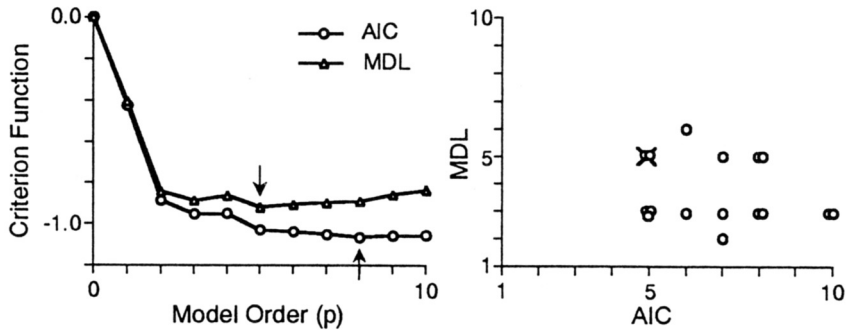


Figure 4.16: In the left panel, example *AIC* and *MDL* criterion functions are shown for the data used in Fig. 4.15. The observations thus consist of two sinusoids and additive first-order lowpass noise. Each frame has a duration of 512 samples and was viewed through a Hanning window. This sum of three autoregressive sequences does *not* have an all-pole spectrum but serves to illustrate the difficulties in applying model-based criteria in the real world. The "correct" model order is five; the example shows *MDL* correctly judging the model order while *AIC*'s estimate is high (eight). The right panel summarizes the two algorithm's estimates for 16 different realizations of the same prototypical signal. The "correct" choice of five is indicated by the X. Although neither algorithm performed well, *AIC*'s estimates are consistently higher than *MDL*'s, which theory predicts.

As p increases, the first term decreases while the second increases, thereby guaranteeing a unique minimum. For the *AR* parameter estimation problem, the negative logarithm of the likelihood function is proportional to the logarithm of the prediction error. Minimizing either of these criteria selects a model order best fitting the observations in terms of spectral variation with frequency and of data length [69].

$$\boxed{\begin{aligned} AIC(p) &= \ln \hat{\sigma}_W^2(p) + \frac{2p}{D_e} \\ MDL(p) &= \ln \hat{\sigma}_W^2(p) + \frac{p \ln D}{D_e} \end{aligned}}$$

The quantity D_e denotes the effective number of observations used in the analysis because of data windowing. It is expressed by the ratio of the energy when the window is applied to that obtained when a rectangular window is used. For example, $D_e = 0.374D$ for the Hanning window. These criterion functions appear to differ little but they do have important differences (Fig. 4.16). The *AIC* estimate is inconsistent [56]: The estimate is usually too large no matter how many observations are available. In contrast, the *MDL* estimate is always consistent [83]. Thus, the *MDL* criterion is more accurate when a "large"* number of observations is available; with fewer observations, either model order method can be used to advantage.

Statistical characteristics of the *AR* spectral estimate are quite different from those of the others described here [6, 15]. Although the probability distribution of the *AR* spectral estimate is not known, the estimate's confidence interval has a F distribution having $(2, D-p)$ degrees of freedom. This result means that the spectral estimate can demonstrate fairly large deviations from the true spectrum, with deviations roughly proportional to the true value.

Using matrix notation, we can express the *AR* spectral estimate by

$$\widehat{\mathscr{S}}_S^{\mathrm{AR}}(f) = \hat{\sigma}_W^2 \left| \mathbf{e}'(f) \widehat{\mathbf{K}}_S^{-1} \boldsymbol{\delta} \right|^{-2}$$

Interestingly, this formula *cannot* be expressed as the expected value of a weighted combination of the observations. The *AR* estimate does have data-dependent spectral resolution, which tends to be superior to the minimum variance estimate in many situations. The estimate also tends to produce ripples in the spectrum,

*What is left unsaid is how large is "large." Despite inconsistency, the larger value yielded by the *AIC* technique is the correct way to err: better too many parameters than not enough.

particularly when the spectrum spans a wide range of values (as when sinusoids are present in the observations). These ripples are akin to sidelobes in window-smoothing functions; because of the data-dependent nature of the estimate, these ripples vary with the observations.

ARMA spectral estimation. Before turning to signal models having both poles and zeros, the simpler case of moving average (MA) spectral estimation should be mentioned. The MA power spectrum estimate has the form

$$\widehat{\mathscr{S}_S}(f) = \hat{\sigma}_W^2 \left|1 + \hat{b}_1 e^{-j2\pi f} + \cdots + \hat{b}_q e^{-j2\pi fq}\right|^2$$

The model parameters need not be found to compute this estimate: Given the first $q+1$ lags of the covariance function, the MA estimate amounts to computing the Fourier Transform of these covariance function values. This estimate is hardly novel and is not frequently referenced. It amounts to applying a much narrower rectangular window than the observation length to the estimated covariance function.

In reality, models for the observations frequently require *both* poles and zeros. For example, consider the simple case in which the observations consist of a sum of signals, each of which has a first-order spectrum. The power spectrum of the sum has *both* poles and zeros, whereas neither component has zeros.

$$\frac{1}{(1-a_1e^{-j2\pi f})(1-a_1e^{+j2\pi f})} + \frac{1}{(1-a_2e^{-j2\pi f})(1-a_2e^{+j2\pi f}) =}$$

$$\frac{\left(1-a_1e^{-j2\pi f}\right)\left(1-a_1e^{+j2\pi f}\right) + \left(1-a_2e^{-j2\pi f}\right)\left(1-a_2e^{+j2\pi f}\right)}{(1-a_1e^{-j2\pi f})(1-a_1e^{+j2\pi f})(1-a_2e^{-j2\pi f})(1-a_2e^{+j2\pi f})}$$

An $ARMA$ model for observations is quite realistic and can encompass a much wider variety of situations than AR ones. Unfortunately, the equations that govern the $ARMA$ model's parameters are nonlinear. In contrast, the equations governing the AR parameters are linear, and the only computational issue is how quickly they can be solved. The issue in $ARMA$ spectral estimation is *how* to solve them. We can derive a set of equations that determine the $ARMA$ parameters by considering the difference equation that governs the observations' covariance function. Using the difference equation for the observations (Eq. 4.48 {124}), we find their covariance function obeys

$$K_S(m) = \begin{cases} \sum_{k=1}^{p} a_k K_S(m-k) + K_{WS}(m) + \sum_{k=1}^{q} b_k K_{WS}(m-k), & m = 0,\ldots,q \\ \sum_{k=1}^{p} a_k K_S(m-k), & m > q \end{cases}$$

$K_{WS}(m-k) = \mathsf{E}[W(l-k)S(l-m)]$ is the cross-covariance function between the observations and the model's input. Because the process $W(\cdot)$ is white, this cross-covariance function equals the model's unit-sample response times the variance of the white-noise input.

$$\begin{aligned} K_{WS}(m-k) &= \mathsf{E}\left[\sum_{n=-\infty}^{l-m} W(l-k)\cdot W(n)h(l-m-n)\right] \\ &= \sigma_W^2 h(k-m) \end{aligned}$$

The equations that must be solved for the $ARMA$ parameters become

$$K_S(m) = \begin{cases} \sum_{k=1}^{p} a_k K_S(m-k) + \sigma_W^2 h(-m) + \sigma_W^2 \sum_{k=1}^{q} b_k h(k-m), & m = 0,\ldots,q \\ \sum_{k=1}^{p} a_k K_S(m-k), & m > q \end{cases}$$

The unit-sample response is, of course, dependent on the values of the parameters a_k and b_k. The multiplication of the unit-sample response and the coefficients b_k is thus nonlinear. Much active research is focused on efficient methods of solving these equations [60: chapter 10];[71: chapter 10].

4.8 Probability Density Estimation

Many signal processing algorithms, implicitly or explicitly, assume that the signal and the observation noise are each well described as Gaussian random sequences. Virtually all linear estimation and prediction filters minimize the mean-squared error while not explicitly assuming any form for the amplitude distribution of the signal or noise. In many formal waveform estimation theories where probability density is, for better or worse, specified, the mean-squared error arises from Gaussian assumptions. A similar situation occurs explicitly in detection theory. The matched filter is provably the optimum detection rule *only* when the observation noise is Gaussian. When the noise is non-Gaussian, the detector assumes some other form. Much of what has been presented in this chapter is based *implicitly* on a Gaussian model for both the signal and the noise. When non-Gaussian distributions are assumed, the quantities upon which optimal linear filtering theory are based, covariance functions, no longer suffice to characterize the observations. While the joint amplitude distribution of any zero-mean, stationary Gaussian stochastic process is entirely characterized by its covariance function; non-Gaussian processes require more. Optimal linear filtering results can be applied in non-Gaussian problems, but we should realize that other informative aspects of the process are being ignored.

This discussion would seem to be leading to a formulation of optimal filtering in a non-Gaussian setting. Would that such theories were easy to use; virtually all of them require knowledge of process characteristics that are difficult to measure and the resulting filters are typically nonlinear [66: chapter 8]. Rather than present preliminary results, we take the tack that knowledge is better than ignorance: At least the first-order amplitude distribution of the observed signals should be considered during the signal processing design. If the signal is found to be Gaussian, then linear filtering results can be applied with the knowledge than no other filtering strategy will yield better results. If non-Gaussian, the linear filtering can still be used and the engineer must be aware that future systems might yield "better" results.*

4.8.1 Types

When the observations are discrete-valued or made so by digital-to-analog converters, estimating the probability mass function is straightforward: Count the relative number of times each value occurs. Let $X(0),\ldots,X(L-1)$ denote a sequence of observations, each of which takes on values from the set $\mathscr{A} = \{a_1,\ldots,a_N\}$. This set is known as an *alphabet* and each a_n is a letter in that alphabet. We estimate the probability that an observation equals one of the letters according to

$$\widehat{P}_X(a_n) = \frac{1}{L}\sum_{l=0}^{L-1} \mathrm{I}\big[X(l) = a_n\big] \,,$$

where $\mathrm{I}[\cdot]$ is the indicator function, equaling one if its argument is true and zero otherwise. This kind of estimate is known in information theory as a *type* [23: Chap. 11], and types have remarkable properties. For example, if the observations are statistically independent, the probability that a given sequence occurs equals

$$\Pr[\mathbf{X} = \{X(0),\ldots,X(L-1)\}] = \prod_{l=0}^{L-1} P_X(X(l)) \,.$$

Evaluating the logarithm, we find that

$$\log\Pr[\mathbf{X}] = \sum_{l=0}^{L-1} \log P_X(X(l))$$

*Note that linear filtering optimizes the mean-squared error whether the signals involved are Gaussian or not. Other error criteria might better capture unexpected changes in signal characteristics and non-Gaussian processes contain internal statistical structure beyond that described by the covariance function.

Note that the number of times each letter occurs equals $L\widehat{P}_X(a_n)$. Using this fact, we can convert this sum to a sum over letters.

$$\begin{aligned}\log \Pr[\mathbf{X}] &= \sum_{n=0}^{N-1} L\widehat{P}_X(a_n) \log P_X(a_n) \\ &= L \sum_{n=0}^{N-1} \widehat{P}_X(a_n) \left[\log \widehat{P}_X(a_n) - \log \frac{\widehat{P}_X(a_n)}{P_X(a_n)} \right] \\ \log \Pr[\mathbf{X}] &= -L \left[\mathscr{H}(\widehat{P}_X) + \mathscr{D}(\widehat{P}_X \| P_X) \right]\end{aligned}$$

which yields

$$\Pr[\mathbf{X}] = \exp\left\{ -L \left[\mathscr{H}(\widehat{P}_X) + \mathscr{D}(\widehat{P}_X \| P_X) \right] \right\} \tag{4.51}$$

We introduce the *entropy* $\mathscr{H}(P)$ [23: §2.1] of a probability distribution and the *Kullback-Leibler distance* $\mathscr{D}(P_1 \| P_0)$ between two distributions {23}.

$$\begin{aligned}\mathscr{H}(P) &= -\sum_{n=0}^{N-1} P(a_n) \log P(a_n) \\ \mathscr{D}(P_1 \| P_0) &= \sum_{n=0}^{N-1} P_1(a_n) \log \frac{P_1(a_n)}{P_0(a_n)}\end{aligned}$$

Clearly, entropy is always non-negative since probabilities are less than or equal to one.* Note that the Kullback-Leibler distance is also non-negative and that P appears in (4.51) in the distance term. Consequently, we maximize Eq. (4.51) with respect to P by choosing $P = \widehat{P}$: *The type estimator is the maximum likelihood estimator of* P_X.

The number of length-L observation sequences having a given type $\widehat{P}$ approximately equals $e^{-L\mathscr{H}(\widehat{P})}$. The probability that a given sequence has a given type approximately equals $e^{-L\mathscr{D}(\widehat{P}\|P)}$, which means that the probability a given sequence has a type *not* equal to the true distribution decays exponentially with the number of observations. Thus, while the coin flip sequences {H,H,H,H,H} and {T,T,H,H,T} are equally likely (assuming a fair coin), the second is more *typical* because its type is closer to the true distribution.

4.8.2 Histogram Estimators

By far the most used technique for estimating the probability distribution of a continuous-valued random variable is the *histogram*; more sophisticated techniques are discussed in [84]. For real-valued data, subdivide the real line into N intervals $(X_i, X_{i+1}]$ having widths $\delta_i = X_{i+1} - X_i$, $i = 1, \ldots, N$. These regions are called "bins" and they should encompass the range of values assumed by the data. For large values, the "edge bins" can extend to infinity to catch the overflows. Given L observations of a stationary random sequence $X(l)$, $l = 0, \ldots, L-1$, the histogram estimate $h(i)$ is formed by simply forming a type from the number L_i of these observations that fall into the i^{th} bin and dividing by the binwidth δ_i.

$$\hat{p}_X(X) = \begin{cases} h(1) = \frac{L_1}{L\delta_1} & X_1 < X \le X_2 \\ h(2) = \frac{L_2}{L\delta_2} & X_2 < X \le X_3 \\ \vdots & \\ h(N) = \frac{L_N}{L\delta_N} & X_N < X \le X_{N+1} \end{cases}$$

The histogram estimate resembles a rectangular approximation to the density. Unless the underlying density has the same form (a rare event), the histogram estimate does *not* converge to the true density as

*If a probability is zero, $0 \log 0$ is defined to be zero.

the number L of observations grows. Presumably, the value of the histogram at each bin converges to the probability that the observations lie in that bin.

$$\lim_{L\to\infty}\frac{L_i}{L}=\int_{X_i}^{X_{i+1}}p_X(X)\,dX$$

To demonstrate this intuitive feeling, we compactly denote the histogram estimate by using the indicator function. This indicator function $\mathrm{I}_i[X(l)]$ equals one if the observation $X(l)$ lies in the i^{th} bin and is zero otherwise. The estimate is simply the average of the indicator functions across the observations.

$$h(i)=\frac{1}{L\delta_i}\sum_{l=0}^{L-1}\mathrm{I}_i[X(l)]$$

The expected value of $\mathrm{I}_i[X(l)]$ is simply the probability P_i that the observation lies in the i^{th} bin. Thus, the expected value of each histogram value equals the integral of the actual density over the bin, showing that the histogram is an unbiased estimate of this integral. Convergence can be tested by computing the variance of the estimate. The variance of one bin in the histogram is given by

$$\mathsf{var}[h(i)]=\frac{P_i-P_i^2}{L\delta_i^2}+\frac{1}{L^2\delta_i^2}\sum_{k\neq l}\left(\mathsf{E}\{\mathrm{I}_i[X(k)]\mathrm{I}_i[X(l)]\}-P_i^2\right)$$

To simplify this expression, the correlation between the observations must be specified. If the values are statistically independent (we have white noise), each term in the sum becomes zero and the variance is given by $\mathsf{var}[h(i)]=(P_i-P_i^2)/(L\delta_i^2)$. Thus, the variance tends to zero as $L\to\infty$ and the histogram estimate is consistent, converging to P_i/δ_i. If the observations are not white, convergence becomes problematical. Assume, for example, that $\mathrm{I}_i[X(k)]$ and $\mathrm{I}_i[X(l)]$ are correlated in a first-order, geometric fashion.

$$\mathsf{E}\{\mathrm{I}_i[X(k)]\mathrm{I}_i[X(l)]\}-P_i^2=P_i^2\rho^{|k-l|}$$

The variance does increase with this presumed correlation until, at the extreme ($\rho=1$), the variance is a constant independent of L! In summary, if the observations are mutually correlated and the histogram estimate converges, the estimate converges to the proper value but more slowly than if the observations were white. The estimate may not converge if the observations are heavily dependent from index to index. This type of dependence structure occurs when the power spectrum of the observations is lowpass with an extremely low cutoff frequency.

Convergence to the density rather than its integral over a region can occur if, as the amount of data grows, we reduce the binwidth δ_i and increase N, the number of bins. However, if we choose the binwidth too small for the amount of available data, few bins contain data and the estimate is inaccurate. Letting X' denote the midpoint of a bin, using a Taylor expansion about this point reveals that the mean-squared error between the histogram and the density at that point is [88: 44–59]

$$\mathsf{E}\left\{[p_X(X')-h(i)]^2\right\}=\frac{p_X(X')}{2L\delta_i}+\frac{\delta_i^4}{36}\left[\left.\frac{d^2p_X(X)}{dX^2}\right|_{X=X'}\right]^2+O\left(\frac{1}{L}\right)+O(\delta_i^5)$$

This mean-squared error becomes zero *only* if $L\to\infty$, $L\delta_i\to\infty$, *and* $\delta_i\to 0$. Thus, the binwidth must decrease *more slowly* than the rate of increase of the number of observations. We find the "optimum" compromise between the decreasing binwidth and the increasing amount of data to be*

$$\delta_i=\left[\frac{9p_X(X')}{2\left[d^2p_X(X)/dX^2|_{X=X'}\right]^2}\right]^{1/5}L^{-1/5}$$

Using this binwidth, we find the the mean-squared error to be proportional to $L^{-4/5}$. We have thus discovered the famous "4/5" rule of density estimation; this is one of the few cases where the variance of a convergent statistic decreases more slowly than the reciprocal of the number of observations. In practice, this optimal binwidth cannot be used because the proportionality constant depends of the unknown density being estimated. Roughly speaking, wider bins should be employed where the density is changing slowly. How the optimal binwidth varies with L can be used to adjust the histogram estimate as more data become available.

*This result assumes that the second derivative of the density is nonzero. If it is not, either the Taylor series expansion brings higher order terms into play or, if all the derivatives are zero, no optimum binwidth can be defined for minimizing the mean-squared error.

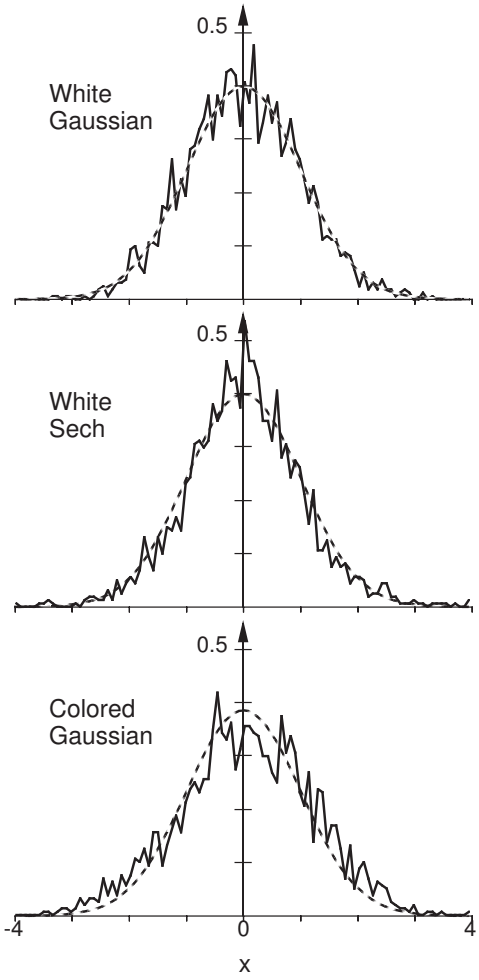


Figure 4.17: Three histogram density estimates are shown and compared with Gaussian densities having the same mean and variance. The histogram on the top is obtained from Gaussian data that are presumed to be white. The middle one is obtained from a non-Gaussian distribution related to the hyperbolic secant $[p_X(X) = \frac{1}{2\sigma}\text{sech}^2(\pi X/2\sigma)]$. This density resembles a Gaussian about the origin but decreases exponentially in the tails. The bottom histogram is taken from a first-order autoregressive Gaussian signal. Thus, these data are correlated, but yield a histogram resembling the true amplitude distribution. In each case, 2000 data points were used and the histogram contained 100 bins.

4.8.3 Density Verification

Once a density estimate is produced, the class of density that best coincides with the estimate remains an issue: Is the density just estimated statistically similar to a Gaussian? The histogram estimate can be used directly in a hypothesis test to determine similarity with any proposed density. Assume that the observations are obtained from a white, stationary, stochastic sequence. Let $\mathscr{M}_0$ denote the hypothesis that the data have an amplitude distribution equal to the presumed density and $\mathscr{M}_1$ the dissimilarity hypothesis. If $\mathscr{M}_0$ is true, the estimate for each bin should not deviate greatly from the probability of a randomly chosen datum lying in the bin. We determine this probability from the presumed density by integrating over the bin. Summing these deviations over the entire estimate, the result should not exceed a threshold. The theory of standard hypothesis testing requires us to produce a specific density for the alternative hypothesis $\mathscr{M}_1$. We cannot

rationally assign such a density; consistency is being tested, not whether either of two densities provides the best fit. However, taking inspiration from the Neyman-Pearson approach to hypothesis testing (§5.1.2 {153}), we can develop a test statistic and require its statistical characteristics *only* under $\mathscr{M}_0$. The typically used, but *ad hoc* test statistic $S(L,N)$ is related to the histogram estimate's mean-squared error [25: 416–41].

$$S(L,N) = \sum_{i=1}^{N} \frac{(L_i - LP_i)^2}{LP_i} = \sum_{i=1}^{N} \frac{L_i^2}{LP_i} - L$$

This statistic sums over the various bins the squared error of the number of observations relative to the expected number. For large L, $S(L,N)$ has a χ^2 probability distribution with $N-1$ degrees of freedom [25: 417]. Thus, for a given number of observations L we establish a threshold η_N by picking a false-alarm probability P_F and using tables to solve $\Pr[\chi^2_{N-1} > \eta_N] = P_F$. To enhance the validity of this approximation, statisticians recommend selecting the binwidth so that each bin contains at least ten observations. In practice, we fulfill this criterion by merging adjacent bins until a sufficient number of observations occur in the new bin and defining its binwidth as the sum of the merged bins' widths. Thus, the number of bins is reduced to some number N', which determines the degrees of freedom in the hypothesis test. The similarity test between the histogram estimate of a probability density function and an assumed ideal form becomes

$$S(L,N') \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \eta_{N'}$$

In many circumstances, the formula for the density is known but not some of its parameters. In the Gaussian case, for example, the mean or variance are usually unknown. These parameters must be determined from the same data used in the consistency test before the test can be used. Doesn't the fact that we use estimates rather than actual values affect the similarity test? The answer is "yes," but in an interesting way: The similarity test changes only in that the number of degrees of freedom of the χ^2 random variable used to establish the threshold is reduced by one for each estimated parameter. If a Gaussian density is being tested, for example, the mean and variance usually need to be found. The threshold should then be determined according to the distribution of a $\chi^2_{N'-3}$ random variable.

Example

Three sets of observations are considered: Two are drawn from a Gaussian distribution and the other not. The first Gaussian example is white noise, a signal whose characteristics match the assumptions of this section. The second is non-Gaussian, which should not pass the test. Finally, the last test consists of colored Gaussian noise that, because of dependent samples, does not have as many degrees of freedom as would be expected. The number of data available in each case is 2000. The histogram estimator uses fixed-width bins and the χ^2 test demands at least ten observations per merged bin. The mean and variance estimates are used in constructing the nominal Gaussian density. The histogram estimates and their approximation by the nominal density whose mean and variance were computed from the data are shown in Fig. 4.17. The chi-squared test ($P_F = 0.1$) yielded the following results.

Density	N'	$\chi^2_{N'-3}$	$S(2000,N')$
White Gaussian	70	82.2	78.4
White sech	65	76.6	232.6
Colored Gaussian	65	76.6	77.8

The white Gaussian noise example clearly passes the χ^2 test. The test correctly evaluated the non-Gaussian example, but declared the colored Gaussian data to be non-Gaussian, yielding a value near the threshold. Failing in the latter case to correctly determine the data's Gaussianity, we see that the χ^2 test is sensitive to the statistical independence of the observations.

Problems

4.1 Estimates of identical parameters are heavily dependent on the assumed underlying probability densities. To understand this sensitivity better, consider the following variety of problems, each of which asks for estimates of quantities related to variance. Determine the bias and consistency in each case.

(a) Compute the maximum *a posteriori* and maximum likelihood estimates of θ based on L statistically independent observations of a Maxwellian random variable X.

$$p_{X|\theta}(X|\theta) = \sqrt{\frac{2}{\pi}}\theta^{-3/2}X^2 e^{-\frac{1}{2}X^2/\theta} \qquad X>0, \theta>0$$
$$p_\theta(\theta) = \lambda e^{-\lambda\theta}, \qquad \theta>0$$

(b) Find the maximum *a posteriori* estimate of the variance σ^2 from L statistically independent observations having the exponential density

$$p_{X|\sigma^2}(X|\sigma^2) = \frac{1}{\sqrt{\sigma^2}} e^{-X/\sqrt{\sigma^2}} \quad X>0$$

where the variance is uniformly distributed over the interval $[0, \sigma^2_{\max})$.

(c) Find the maximum likelihood estimate of the variance of L identically distributed, but dependent Gaussian random variables. Here, the covariance matrix is written $\mathbf{K}_X = \sigma^2\widetilde{\mathbf{K}}_X$, where the normalized covariance matrix has trace $\mathrm{tr}[\widetilde{\mathbf{K}}_X] = L$. Assume the random variables have zero mean.

4.2 Imagine yourself idly standing on the corner in a large city when you note the serial number of a passing beer truck. Because you are idle, you wish to estimate (guess may be more accurate here) how many beer trucks the city has from this single observation.

(a) Making appropriate assumptions, the beer truck's number is drawn from a uniform probability density ranging between zero and some unknown upper limit, find the maximum likelihood estimate of the upper limit.

(b) Show that this estimate is biased.

(c) In one of your extraordinarily idle moments, you observe throughout the city L beer trucks. Assuming them to be independent observations, now what is the maximum likelihood estimate of the total?

(d) Is this estimate of θ biased? asymptotically biased? consistent?

4.3 Faulty Geiger Counter
A Geiger counter detects the incidence of radioactive particles by counting the number hitting the detector over each one second interval. Our faulty detector only counts each particle with probability p. The physics of particle generation has the particles emitted according to a Poisson process, making the number emitted each second governed by a Poisson distribution.

$$\Pr[N=n] = \frac{\lambda^n e^{-\lambda}}{n!}$$

(a) What is the probability distribution of number of particles the faulty Geiger actually detects in one second?

(b) Find the minimum-mean-squared-error estimate of N based on observing the Geiger counter output.

(c) Examine the estimator's bias and variance characteristics.

4.4 Estimating Power
X_t is a wide-sense stationary, Gaussian process having mean m_X and correlation function $R_X(\tau)$. The mean is known; the correlation function is also known up to a scale factor. We want to estimate the power P_X in this process: $P_X = R_X(0)$. Toward this end, the process is sampled (sampling interval T), providing L samples X_{lT}, $l = 0, \ldots, L-1$.

(a) One way to estimate the power is to compute the average of each sample's power.

$$\widehat{P_X} = \frac{1}{L}\sum_{l=0}^{L-1} X_{lT}^2$$

Is this estimate biased? Indicate your reasoning.

(b) Under what conditions is this estimate consistent?

(c) Another, better justified approach, is to consider P_X a parameter of the probability distribution governing the observations and use a formal parameter estimation technique. Find an estimate for P_X using this approach.

4.5 Estimating a Bit

To send a bit, a discrete-time communications system transmits either $+1$ or -1 for L successive indices. The channel adds white Gaussian noise and the receiver must determine which bit was sent from these noise-corrupted observations. The bit's values are equally likely.

(a) What is the MAP estimate of the bit's value?

(b) Determine the bias, if any, of the MAP estimate?

(c) Is the MAP estimate in this case consistent?

(d) Find the minimum mean-squared error estimator of the bit.

4.6 Sequential Estimation

We make L observations $X_1,\ldots,X_L$ of a parameter θ corrupted by additive noise ($X_l = \theta + N_l$). The parameter θ is a Gaussian random variable [$\theta \sim \mathcal{N}(0,\sigma_\theta^2)$] and N_l are statistically independent Gaussian random variables [$N_l \sim \mathcal{N}(0,\sigma_N^2)$].

(a) Find the *MMSE* estimate of θ.

(b) Find the maximum *a posteriori* estimate of θ.

(c) Compute the resulting mean-squared error for each estimate.

(d) Consider an alternate procedure based on the same observations X_l. Using the *MMSE* criterion, we estimate θ immediately after each observation. This procedure yields the sequence of estimates $\widehat{\theta}_1(X_1), \widehat{\theta}_2(X_1,X_2),\ldots,\widehat{\theta}_L(X_1,\ldots,X_L)$. Express $\widehat{\theta}_l$ as a function of $\widehat{\theta}_{l-1}$, σ_{l-1}^2, and X_l. Here, σ_l^2 denotes the mean-squared estimation error of the l^{th} estimate. Show that

$$\frac{1}{\sigma_l^2} = \frac{1}{\sigma_\theta^2} + \frac{l}{\sigma_N^2}$$

4.7 Phase Estimation

A sinusoid of known amplitude and frequency but unknown phase is observed in zero-mean white Gaussian noise.

$$X_l = A\cos(2\pi f_0 l + \theta) + N_l,\ 0 \le l \le L-1$$

Assume the observation interval contains an integer number of the sinusoid's periods.

(a) What is the maximum likelihood estimate of the phase?

(b) Is your estimate biased?

(c) What is the Cramér-Rao bound on the mean-squared estimation error?

4.8 Estimating Phase, Amplitude and Frequency

You are given the discrete-time signal $A\cos(2\pi f_0 l - \theta)$ observed in the presence of white Gaussian noise having zero-mean and variance σ^2. Eventually, we want to find all the parameters, but let's build up to that by estimating first the phase θ, then the amplitude A, finally incorporating the frequency f_0. Throughout assume the number of observations L contains an integer number of periods of the sinusoidal signal.

(a) What is the maximum likelihood estimate for the phase assuming the amplitude and frequency are known?

(b) Find the Cramér-Rao bound for your estimate.

(c) Create a MATLAB simulation of your estimation procedure. Let $A = 1$ and $f_0 = 100/L$, with $L = 1024$. Run 1,000-trial simulations to estimate the phase for $\theta = \pi/4$ for signal-to-noise ratios A^2/σ^2 of 1.0 and 0.04. Calculate the empirical mean and standard deviation of your estimates. Do they agree with theory?

(d) What are the maximum likelihood estimates of the phase and the amplitude? Find the Cramér-Rao bound for this case.

(e) Using the same simulations as in part (c), do these estimates have the predicted statistics?

(f) Find the joint maximum likelihood estimates for all three parameters. Calculate the Cramér-Rao lower bound for each parameter's estimation error.

Note: The analytic maximum likelihood estimate of the frequency is difficult to apply to data. Find an equivalent empirical estimate.

(g) For the data file `sineinnoise.mat`, estimate the amplitude, frequency and phase. Indicate the possible range of values for your estimates (*i.e.*, provide error bars).

4.9 Gauss and Maximum Likelihood

Although the maximum likelihood estimation procedure was not clearly defined until early in the 20th century, Gauss showed in 1805 that the Gaussian density* was the *sole* density for which the maximum likelihood estimate of the mean equaled the sample average. Let $\{X_0, \ldots, X_{L-1}\}$ be a sequence of statistically independent, identically distributed random variables.

(a) What equation defines the maximum likelihood estimate $\widehat{m}_{\mathrm{ML}}$ of the mean m when the common probability density function of the data has the form $p(X-m)$?

(b) The sample average is, of course, $\sum_l X_l/L$. Show that it minimizes the mean-squared error $\sum_l [X_l - m]^2$.

(c) Equating the sample average to $\widehat{m}_{\mathrm{ML}}$, combine this equation with the maximum likelihood equation to show that the Gaussian density uniquely satisfies the equations.

Note: Because both equations equal 0, they can be equated. Use the fact that they must hold for *all* L to derive the result. Gauss thus showed that mean-squared error and the Gaussian density were closely linked, presaging ideas from modern robust estimation theory.

4.10 A Non-Gaussian Problem

We observe a sequence of L statistically independent *Laplacian* random variables having an unknown mean and variance.

(a) Find the maximum likelihood estimates of the mean and variance.

(b) Is the estimate of the mean biased?

(c) What are the smallest mean-squared errors that these estimates can achieve?

4.11 What's In-Between the Samples?

We sample a stationary random process X_t every T seconds, ignoring whether the process is bandlimited or not. To reconstruct the signal from the samples, we use *linear interpolation*.

$$\widehat{X}_t = aX_{(n-1)T} + bX_{nT}, \quad (n-1)T \le t \le nT$$

(a) Find the minimum mean-squared error linear interpolator. In other words, what are the best values for a and b?

(b) Show that the maximum likelihood interpolator is also linear when X_t is a wide-sense stationary, zero-mean, Gaussian process. In other words, if $X_{(n-1)T}$ and X_{nT} comprise your observations, show that the maximum likelihood estimate of X_t has the linear form given above. For this part, you do not need to find a and b.

*It wasn't called the Gaussian density in 1805; this result is one of the reasons why it is.

(c) Find the Cramér-Rao bound for the interpolation estimate of X_t.

4.12 Maximum Likelihood Estimation of a Function of a Parameter

In an example {69}, we derived the maximum likelihood estimate of the mean and variance of a Gaussian random vector. You might wonder why we chose to estimate the variance σ^2 rather than the standard deviation σ. Using the same assumptions provided in the example, let's explore the consequences of estimating a *function* of a parameter [89: Probs. 2.4.9, 2.4.10].

(a) Assuming that the mean is known, find the maximum likelihood estimates of first the variance, then the standard deviation.

(b) Are these estimates biased?

(c) Describe how these two estimates are related. Assuming that $f(\cdot)$ is a monotonic function, how are $\widehat{\theta}_{\mathrm{ML}}$ and $\widehat{f(\theta)}_{\mathrm{ML}}$ related in general?

These results suggest a general question. Consider the problem of estimating some function of a parameter θ, say $f_1(\theta)$. The observed quantity is X and the conditional density $p_{X|\theta}(X|\theta)$ is known. Assume that θ is a nonrandom parameter.

(d) What are the conditions for an efficient estimate $\widehat{f_1(\theta)}$ to exist?

(e) What is the lower bound on the variance of the error of any unbiased estimate of $f_1(\theta)$?

(f) Assume an efficient estimate of $f_1(\theta)$ exists; when can an efficient estimate of some other function $f_2(\theta)$ exist?

4.13 Imposing a Model on the Data

We have an IID sequence of Gaussian random vectors $\mathbf{X}_l$, $l = 0, \ldots, L-1$, each of which has dimension N. Each $\mathbf{X}_l$ has a Gaussian probability distribution, with mean $\mathbf{m}$ and covariance matrix $\sigma^2\mathbf{I}$. We want to estimate the 'strength" and direction of the mean. In other words, we want to describe the mean as $\mathbf{m} = \beta\mathbf{u}$, where β is some scalar and $\mathbf{u}$ has unit norm.

(a) Assuming the variance σ^2 is known, find the maximum likelihood estimates of β and $\mathbf{u}$.

(b) Find the Fisher information matrix for these parameters. Interpret the structure of this matrix.

(c) Suppose the variance is not known. What are the maximum likelihood estimates of β, $\mathbf{u}$ and σ^2?

4.14 Optimality of Maximum Likelihood Estimates

Let the observations $X(l)$ consist of statistically independent, identically distributed Gaussian random variables having zero mean but unknown variance. We wish to estimate σ^2, their variance.

(a) Find the maximum likelihood estimate $\widehat{\sigma^2}_{\mathrm{ML}}$ and compute the resulting mean-squared error.

(b) Show that this estimate is efficient.

(c) Consider a new estimate $\widehat{\sigma^2}_{\mathrm{NEW}}$ given by $\widehat{\sigma^2}_{\mathrm{NEW}} = \alpha\widehat{\sigma^2}_{\mathrm{ML}}$, where α is a constant. Find the value of α that minimizes the mean-squared error for $\widehat{\sigma^2}_{\mathrm{NEW}}$. Show that the mean-squared error of $\widehat{\sigma^2}_{\mathrm{NEW}}$ is less than that of $\widehat{\sigma^2}_{\mathrm{ML}}$. Is this result compatible with part b?

4.15 Method of Moments

An early twentieth-century estimation approach used sample moments—sample average, average mean-square value, etc.—as a basis for estimating appropriate quantities. While its properties are difficult to determine, it can yield a simpler estimate than "standard," more modern methods. For example, consider a wide-sense stationary, white process having an amplitude density given by a so-called *mixture* distribution.

$$p_X(X) = ap_1(X) + (1-a)p_2(X)$$

Here, a is known as the mixture parameter and $p_1(X)$, $p_2(X)$ are two known probability densities. We want to estimate the mixture parameter for the special case wherein the two densities are zero-mean Gaussians having variances σ_1^2, σ_2^2 respectively. We have L statistically independent observations.

(a) What is the expected value of sample values of the first two moments: the sample average $\overline{X}$ and sample mean-squared value $\overline{X^2}$?

$$\overline{X} = \frac{1}{L}\sum_{l=0}^{L-1} X_l \qquad \overline{X^2} = \frac{1}{L}\sum_{l=0}^{L-1} X_l^2$$

(b) Use your expected value expressions to derive an estimate for the mixture parameter.

(c) Is this estimate biased? consistent? Why or why not?

(d) Find the expected mean-squared error.

4.16 Optimal and Simple Communications

A multiplexed communication system needs to be designed that sends two numbers simultaneously. Perhaps the simplest design represents the numbers as the amplitudes of two carrier signals. The received signal has the form

$$R_l = A_1 c_1(l) + A_2 c_2(l) + N_l, \quad l = 0, \ldots, L-1$$

where N_l is ubiquitous additive (not necessarily white) Gaussian noise. The carrier signals $c_1(l)$ and $c_2(l)$ have unit energy; their detailed waveforms need to be selected to provide the best possible system design.

(a) What is the maximum likelihood estimate of the amplitudes?

(b) Is the maximum likelihood estimate biased or not? If it is biased, what are the most general conditions on the carrier signals and the noise would it make it unbiased?

(c) Under what conditions are the amplitude estimation errors uncorrelated and as small as possible?

4.17 MIMO Channels

Two parameters θ_1, θ_2 are transmitted over a MIMO (Multiple-Input, Multiple-Output) channel. The two parameters constitute the channel's two-dimensional vector input $\boldsymbol{\theta}$, and the channel output is $\mathbf{H}\boldsymbol{\theta}$. $\mathbf{H}$ is the *non-square* "transfer function" matrix that represents the set of linear combinations of the parameters found in the output. The observations consist of

$$\mathbf{R} = \mathbf{H}\boldsymbol{\theta} + \mathbf{N},$$

where the noise vector $\mathbf{N}$ is Gaussian, having zero mean and covariance matrix $\mathbf{K}$.

(a) What is the maximum likelihood estimate of $\boldsymbol{\theta}$?

(b) Find this estimate's total mean-squared error.

(c) Is this estimate biased? Is it efficient?

4.18 Prediction

A signal $s(l)$ can be described as a stochastic process that has zero mean and covariance function $K_s(\ell) = \sigma_s^2 a^{|\ell|}$. This signal is observed in additive white Gaussian noise having variance σ^2. The signal and noise are statistically independent of each other.

(a) Find the optimal predictor $\widehat{s}(l+1)$ that is based on observations that end at time l and begin at time $l-L+1$.

(b) How does this predictor change if we want to estimate $s(l+k)$ based on observations made over $[l, \ldots, l+L-1]$?

(c) How does the predictor's mean-squared error vary with k?

4.19 Let the observations be of the form $\mathbf{X} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}$ where $\boldsymbol{\theta}$ and $\mathbf{n}$ are statistically independent Gaussian random vectors.

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_\theta) \qquad \mathbf{n} \sim \mathcal{N}(\mathbf{0},\ \mathbf{K}_n)$$

The vector $\boldsymbol{\theta}$ has dimension M; the vectors $\mathbf{X}$ and $\mathbf{n}$ have dimension N.

(a) Derive the minimum mean-squared error estimate of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_{\text{MMSE}}$, from the relationship $\widehat{\boldsymbol{\theta}}_{\text{MMSE}} = \mathsf{E}[\boldsymbol{\theta}|\mathbf{X}]$.

(b) Show that this estimate and the optimum linear estimate $\widehat{\boldsymbol{\theta}}_{\text{LIN}}$ derived by the Orthogonality Principle are equal.

(c) Find an expression for the mean-squared error when these estimates are used.

4.20 Suppose we consider an estimate of the parameter θ having the form $\widehat{\theta} = \mathscr{L}(\mathbf{X}) + C$, where $\mathbf{X}$ denotes the vector of the observables and $\mathscr{L}(\cdot)$ is a linear operator. The quantity C is a constant. This estimate is *not* a linear function of the observables unless $C = 0$. We are interested in finding applications for which it is advantageous to allow $C \neq 0$. Estimates of this form we term "quasi-linear."

(a) Show that the optimum (minimum mean-squared error) quasi-linear estimate satisfies

$$\mathsf{E}[\langle \mathscr{L}_\diamond(\mathbf{X}) + C_\diamond - \theta, \mathscr{L}(\mathbf{X}) + C \rangle] = 0, \quad \text{for all } \mathscr{L}(\cdot) \text{ and } C$$

where $\widehat{\theta}_{\text{QLIN}} = \mathscr{L}_\diamond(\mathbf{X}) + C_\diamond$.

(b) Find a general expression for the mean-squared error incurred by the optimum quasi-linear estimate.

(c) Such estimates yield a smaller mean-squared error when the parameter θ has a nonzero mean. Let θ be a scalar parameter with mean m. The observables comprise a vector $\mathbf{X}$ having components given by $X_l = \theta + N_l, l = 1, \ldots, L$ where N_l are statistically independent Gaussian random variables [$N_l \sim \mathscr{N}(0, \sigma_N^2)$] independent of θ. Compute expressions for $\widehat{\theta}_{\text{QLIN}}$ and $\widehat{\theta}_{\text{LIN}}$. Verify that $\widehat{\theta}_{\text{QLIN}}$ yields a smaller mean-squared error when $m \neq 0$.

4.21 On Page 80, we questioned the existence of an efficient estimator for signal parameters. We found in the succeeding example that an unbiased efficient estimator exists for the signal amplitude. Can a nonlinearly represented parameter, such as time delay, have an efficient estimator?

(a) Simplify the condition for the existence of an efficient estimator by assuming it to be unbiased. Note carefully the dimensions of the matrices involved.

(b) Show that the only solution in this case occurs when the signal depends "linearly" on the parameter vector.

4.22 Cramér-Rao Bound for Signal Parameters
In many problems, the signal as well as the noise are sometimes modeled as Gaussian processes. Let's explore what differences arise in the Cramér-Rao bounds for the stochastic and deterministic signal cases. Assume that the signal contains unknown parameters $\boldsymbol{\theta}$, that it is statistically independent of the noise, and that the noise covariance matrix is known.

(a) What forms do the conditional densities of the observations take under the two assumptions? What are the two covariance matrices?

(b) As a preliminary, show that

$$\frac{\partial \mathbf{A}^{-1}}{\partial \theta} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \theta} \mathbf{A}^{-1} .$$

(c) Assuming the stochastic signal model, show that each element of the Fisher information matrix has the form

$$F_{ij} = \frac{1}{2} \operatorname{tr} \left[\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_i} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j} \right]$$

where $\mathbf{K}$ denotes the covariance matrix of the observations. Specialize this expression by assuming the noise component has no unknown parameters.

(d) Compare the stochastic and deterministic bounds, the latter is given by Eq. 4.15 {80}, when the unknown signal parameter is the amplitude. Assume the noise covariance matrix equals $\sigma_N^2 \mathbf{I}$. Do these bounds have similar dependence on signal-to-noise ratio?

4.23 Estimating the Amplitude of a Sinusoid

Suppose you observe a discrete-time sinusoid in additive *Laplacian* white noise having variance per sample of σ^2.

$$X_l = A\sin(2\pi f_0 l) + N_l,\ l = 0,\ldots,L-1$$

The frequency is known and is harmonic with the observation interval ($f_0 = n/L$ for some integer n).

(a) What equation determines the maximum likelihood amplitude estimate?

(b) Because no closed form solution for the estimate is evident, write a MATLAB program that simulates the observations and finds the estimate. Set $L = 1024$ and $f_0 = 100/L$. Let $A = 1$ and $\sigma^2 = 1$. Compute and plot the derivative of the log likelihood function for values of A close to the true amplitude. What do you conclude from this result?

(c) For your dataset, find the maximum likelihood estimate of A.

(d) Find the Cramér-Rao bound for the error variance.

(e) In practice, using Gaussian-noise assumption provides far simpler answers than using a model more closely describing the data. Simulate $1,000$ observational trials using the Laplacian noise description. Use two values for the noise variance — 1 and 25 — in each set of trials. Determine the mean and variance of the Gaussian-derived amplitude estimate. What is the theoretical value for the mean-squared error when the Gaussian-based estimate for the amplitude is used? Does the simulated variance agree with this result?

4.24 BLUE

In addition to forcing a minimum-mean-squared-error estimator to be a linear function of the data, we can also force it to be unbiased. Enforcing these requirements on the estimator makes it the *B*est *L*inear *U*nbiased *E*stimator (BLUE). To explore the properties of this estimator, we consider the problem of estimating the signal amplitude A when we observe

$$X_l = As_l + N_l, \quad 0 \le l < L\,.$$

The signal s_l is known and the noise N_l is wide-sense stationary having covariance function $\mathbf{K}_N(\ell)$.

(a) Find the BLUE estimator for the signal amplitude.

(b) What is an expression for the resultant mean-squared error?

(c) What signal choice minimizes the mean-squared error?

4.25 In Poisson problems, the number of events n occurring in the interval $[0,T)$ is governed by the probability distribution {16}

$$\Pr[n] = \frac{(\lambda_0 T)^n}{n!} e^{-\lambda_0 T}\,,$$

where λ_0 is the average rate at which events occur.

(a) What is the maximum likelihood estimate of average rate?

(b) Does this estimate satisfy the Cramér-Rao bound?

(c) Now suppose the rate varies sinusoidally according to

$$\lambda(t) = \lambda_0 \exp\{a\cos 2\pi f_0 t\} \quad 0 \le t < T$$

where the frequency f_0 is a harmonic of $1/T$. What are the maximum likelihood estimates of λ_0 and a in this case?

Note: The following facts will prove useful.

$I_0(a) = \dfrac{1}{2\pi}\displaystyle\int_0^{2\pi} e^{a\cos\theta}\,d\theta$ is the modified Bessel function of the first kind, order 0.

$I_0'(a) = I_1(a)$, the modified Bessel function of the first kind, order 1.

$I_0''(a) = \dfrac{I_2(a) + I_0(a)}{2}$

(d) Find the Cramér-Rao bounds for the mean-squared estimation errors for $\widehat{\lambda}_0$ and $\widehat{a}$ assuming unbiased estimators.

4.26 In the "classic" radar problem, not only is the time of arrival of the radar pulse unknown but also the amplitude. In this problem, we seek methods of simultaneously estimating these parameters. The received signal $X(l)$ is of the form

$$X(l) = \theta_1 s(l-\theta_2) + N(l)$$

where θ_1 is Gaussian with zero mean and variance σ_1^2 and θ_2 is uniformly distributed over the observation interval. Assuming the noise is white and Gaussian, find the receiver that computes the maximum *a posteriori* estimates of θ_1 and θ_2 jointly. Draw a block diagram of this receiver and interpret its structure.

4.27 We stated without derivation the Cramér-Rao bound for estimates of signal delay (Eq. 4.16 {82}).

(a) The parameter θ is the delay of the signal $s(\cdot)$ observed in additive, white Gaussian noise: $X(l) = s(l-\theta) + N(l)$, $l = 0, \ldots, L-1$. Derive the Cramér-Rao bound for this problem.

(b) On Page 68, this bound is claimed to be given by $\sigma_N^2 / E\beta^2$, where β^2 is the mean-squared bandwidth {68}. Derive this result from your general formula. Does the bound make sense for all values of signal-to-noise ratio E/σ_N^2?

(c) Using optimal detection theory, derive the expression {67} for the probability of error incurred when trying to distinguish between a delay of τ and a delay of $\tau + \Delta$. Consistent with the problem posed for the Cramér-Rao bound, assume the delayed signals are observed in additive, white Gaussian noise.

4.28 System Identification

The system identification problem for linear systems is to determine the system's unit-sample response by applying an input to the system and measuring the resulting output. Normally, finding the unit-sample response would be solved using Fourier transform techniques. However, if additive noise corrupts the observed output, that simple approach won't work. For this problem, assume the system is FIR and its unit-sample response is $h(0), \ldots, h(p-1)$. The input $u(l)$ is known (determined by you!) and is provided for $l = 0, \ldots, L-1$. The output of the system is measured in the presence of additive white Gaussian noise while the input is being provided.

$$X_l = \sum_{k=0}^{p-1} h(k)u(l-k) + N_l,\ l = 0, \ldots, L-1$$

(a) Find the minimum mean-squared error estimate of the unit-sample response.

(b) Is this estimate biased?

(c) Find the Cramér-Rao bound for this problem. How tight is the bound to the actual mean-squared error?

4.29 Estimating Filter Characteristics

White, discrete-time noise is passed through a first-order IIR filter to produce the observations.

$$X_l = aX_{l-1} + W_l, \quad 0 \le l < L$$

Here, W_l is white Gaussian noise, each sample of which has variance σ^2. Both the filter parameter a and the input power σ^2 (the variance of each W_l) are to be estimated.

(a) Find the maximum likelihood estimates of these parameters.

(b) Find the Cramér-Rao bound for any unbiased estimator for this estimation problem.

4.30 Estimating Model Probabilities

We want to estimate the *a priori* probability π_0 based on data obtained over N statistically independent observation intervals. During each length-L observation interval, the observations consist of white Gaussian noise having variance of either σ_0^2 or σ_1^2 ($\sigma_1^2 > \sigma_0^2$). π_0 is the probability that the observations have variance σ_0^2 and we do not know which model applies for any observation interval.

(a) One approach is to classify each observation interval according to its variance, count the number of times the variance equals σ_0^2, and divide this count by N. What is the classification algorithm that lies at the heart of this estimator?

(b) What classifier threshold yields an unbiased estimate of π_0? Comment on the feasibility of this approach.

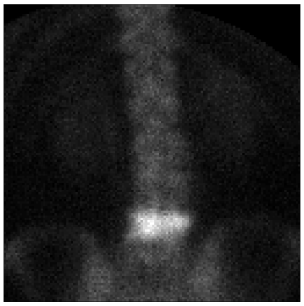
(c) Rather than use this *ad hoc* approach, let's use a systematic estimation approach: what is the maximum likelihood estimate of π_0 based on the N observation intervals?

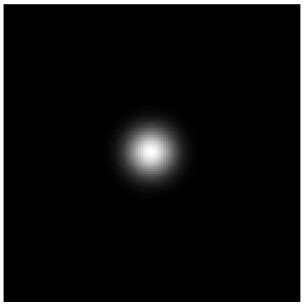
4.31 Image Restoration

Nuclear medicine imaging systems can be modeled as a blurring process (lowpass filtering) plus additive noise.

$$x(m,n) = s(m,n) \otimes g(m,n) + w(m,n)$$

with $g(m,n)$ the unit-sample response of the blurring process, $s(m,n)$ the true image, and $w(m,n)$ white noise. A sample image (`spine.mat`) and the DFT of $g(m,n)$ (`G.mat`) are shown.





The spine image is blurred and noisy and the blurring system is indeed lowpass.

(a) Assume that the power density function of medical images has the form

$$\mathscr{S}_s\left(e^{j2\pi f_1}, e^{j2\pi f_2}\right) = \frac{\sigma_s^2}{\left(1 + k \cdot (2\pi f_1)^2 + k \cdot (2\pi f_2)^2\right)^{\alpha}}, \; -\frac{1}{2} \le f_1, f_2 \le \frac{1}{2}$$

where $\frac{1}{2} \le \alpha \le \frac{3}{2}$, σ_s^2 is the DC power, and k is a constant related to the spatial sampling interval. By examining the power spectrum of the spine image, determine reasonable values for σ_s^2, k, and α.

(b) What is the optimal mean-squared error transfer function? Express it in terms of the ratio $R = \sigma_s^2/\sigma_w^2$.

(c) Implement the optimal restoration filter using MATLAB. Experiment with various choices for α and R, and pick the best restored image you found.

4.32 The signal has a power spectrum given by

$$\mathscr{S}_s(f) = \frac{17}{20} \cdot \frac{1 + \frac{8}{17}\cos 2\pi f}{1 - \frac{4}{5}\cos 2\pi f}$$

This signal is observed in additive white noise having variance equaling 6.

(a) What is the transfer function of the signal's "generation" system? In other words, what system having white noise of spectral height one as its input could have generated the signal?

(b) Find the unit-sample response of the noncausal Wiener filter.

(c) Find the difference equation governing the causal Wiener filter ($l_e = 0$).

4.33 The *Eckhart filter* is an optimum linear filter that maximizes the signal-to-noise ratio of its output [31]. To find the unit-sample response of the *FIR* Eckhart filter, consider observations of the form $\mathbf{X} = \mathbf{s} + \mathbf{N}$ where the covariance matrix of the noise is known. The signal-to-noise ratio is computed according to $\mathsf{E}[\|\mathbf{h}^t\mathbf{s}\|^2] / \mathsf{E}[\|\mathbf{h}^t\mathbf{N}\|^2]$, where $\mathbf{h}$ is the desired unit-sample response.

(a) Assuming the signal is nonrandom, find the Eckhart filter's unit-sample response.

(b) What is the signal-to-noise ratio produced by the Eckhart filter? How does it compare with that produced by the corresponding Wiener filter?

(c) Now assume the signal is random, having covariance matrix $\mathbf{K}_s$. Characterize the Eckhart filter.

4.34 Minimum-Output-Power Filtering

We can design filters for stochastic observations by minimizing the filter's output power while constraining the filter to have certain characteristics. Assume the filter has finite duration L and is to filter a zero-mean random process X_l.

(a) What filter minimizes its output power while having unity gain at zero frequency?

(b) Now suppose the filter must have a unit-sample response that has a mean-square value of one. What is the unit-sample response of this filter? Is it unique?

(c) Suppose the unit-sample response is restricted to be a weighted sum of K signals, $K < L$, where the weights are to be adapted according to filter design criteria. This requirement is most conveniently expressed with a *projection matrix* $\mathbf{P}$: $\mathbf{Ph} = \mathbf{h}$. Projection matrices have eigenvalues equal to one or zero, with the eigenvectors corresponding to unity eigenvalues spanning the subspace to which we want to restrict $\mathbf{h}$ (the unit-sample response of the filter written as a vector). What filter minimizes its output power subject to this constraint?

4.35 Linear Phase Predictors

The standard one-step prediction problem uses an *FIR* predictor of the form

$$\widehat{S}(l) = h_1 s(l-1) + h_2 s(l-2) + \cdots + h_p s(l-p)$$

The optimal coefficients are found from the Yule-Walker equations. Suppose we impose the constraint that the predictor, considered as a *FIR* linear filter, have a linear phase. This additional requirement means that the filter coefficients obey the symmetry condition $h_k = h_{p-k+1}$, $k = 1, \ldots, p/2$. Let the filter's order, p, be an even integer.

(a) What are the system of equations the optimal linear phase predictor must satisfy?

(b) What equation governs the linear-phase *LMS* adaptive filter?

(c) Find conditions on the *LMS* filter's adaptation coefficient μ that guarantee convergence.

4.36 We frequently employ matrix eigenanalysis to explore the convergence of dynamic adaptive methods. Consider the vector difference equation

$$\mathbf{v}(l) = \mathbf{A}\mathbf{v}(l-1) + \mathbf{u}(l)$$

Assume the matrix $\mathbf{A}$ has distinct eigenvalues.

(a) Show that if you define a new state vector to be $\mathbf{x}(l) = \mathbf{M}\mathbf{v}(l)$, where the matrix $\mathbf{M}$ is a unitary matrix, then the new state-transition matrix has the same eigenvalues as $\mathbf{A}$.

(b) There exists a choice for $\mathbf{M}$ that diagonalizes the state-transition matrix. What is the resulting state-transition matrix? For this new state equation, under what conditions is the system stable? Argue this condition applies to the original state equation.

(c) Suppose the input is a unit sample, which evokes a transient response from the system. When the system is stable, the output eventually decays to zero. How quickly does this decay occur?

4.37 On Page 94, the *settling time* of *LMS* filters was related to the eigenvalues of the observations' covariance matrix. A little analysis coupled with engineering knowledge makes these claims clear. Let's discuss the vector difference equation $\mathbf{v}(l) = \mathbf{A}\mathbf{v}(l-1)$, with $\mathbf{A}$ a symmetric matrix.

(a) Because this vector difference equation is first order, it represents a set of coupled first-order difference equations. Thus, some homogeneous M-dimensional linear system is described. Show that the homogeneous response can be expressed as a weighted linear combination of the eigenvectors of $\mathbf{A}$.

(b) Use the matrix's eigen decomposition to show that the norm of $\mathbf{v}$ remains bounded for any initial condition if $\max|\lambda_{\mathbf{A}}| < 1$. Letting $\mathbf{A} = \mathbf{I} - 2\mu\mathbf{K}_X$, show that this bound implies the convergence condition $0 < \mu < 1/\max\lambda_{\mathbf{K}_X}$ for the *LMS* algorithm.

(c) The rate at which a linear system reaches steady state is determined by the mode having the largest time constant. Use this analogy to show that the decay rate of our difference equation is determined by the eigenvalue of $\mathbf{A}$ having the largest magnitude and that the corresponding time constant equals

$$\tau = -\frac{1}{\ln\max|\lambda_{\mathbf{A}}|}$$

(d) Find the convergence rate of the *LMS* adaptive filter that reaches steady state smoothly (no oscillatory adaptation).

4.38 LMS Prediction

We wish to find the one-step predictor for a second-order *AR* process using the *LMS* algorithm. Here, the signal $s(l)$ is governed by the all-pole difference equation

$$s(l) = a_1 s(l-1) + a_2 s(l-2) + w(l)\ ,$$

where $w(l)$ is white Gaussian noise. The signal is observed directly; we seek to predict what its next value will be at each moment of time. Let the *LMS* filter's order be two for the special case of $a_1 = 1.2728$, $a_2 = -.95$, and $\sigma_w^2 = 1$.

(a) What is the theoretical worst-case mean-squared prediction error and the possible prediction error produced by a causal Wiener filter. Express your answer in general terms of the signal generation filter's unit sample response.

(b) Compute the covariance matrix relevant to determining the adaptation parameter μ of the *LMS* algorithm. What is the largest value this parameter can have?

(c) Simulate this *LMS* prediction filter for several values of μ. Show that larger values lead to quicker convergence. Does the "steady-state" mean-square prediction error depend on μ? How does this mean-square prediction error produced by the *LMS* algorithm compare with the theoretical worst- and best-case bounds?

4.39 Adaptive Filtering for Communications

In digital communication, the signals frequently used to represent digital data are square waves. To improve the signal-to-noise ratio, adaptive filters tailored to this problem are used. The filtered result $\widehat{s}(l)$ equals $\mathbf{h}'(l)\mathbf{R}$, with the filter's unit sample response $\mathbf{h}(l)$ updated at each index by the LMS algorithm. As in many adaptive filtering problems, calculating the error signal is both essential to the LMS algorithm and difficult to determine, the difficulty here being we don't know which square wave was sent. Here, the error criterion is adjusted to take into account the signal special property, namely at every sample, the signal's magnitude is a constant (normalized to be unity).

(a) Derive the adaptive filtering algorithm that minimizes the expected squared error, the error given by $\varepsilon(l) = |\widehat{s}(l)|^2 - 1$.

(b) Show that the resulting algorithm will not yield accurate signal estimates. Indicate why the algorithm is flawed.

(c) Change the definition of the error to fix this flaw and derive a better adaptive filtering algorithm.

4.40 Adaptive Line Enhancement

Adaptive line enhancement illustrates well an important application of dynamic adaptive filters. Assume the observations are known to contain a signal (as well as ever-present noise) having period P. This signal is not known more precisely than this: It could be a sinusoid of arbitrary amplitude and phase, or a distorted sinusoid having unknown harmonic structure. We want to design a *FIR* dynamic adaptive filter that enhances the periodic signal (which has a line spectrum) and thereby reduce the observation noise.

(a) The crucial part of the design is the selection of the error signal, whose energy is minimized by the filter. Assuming the noise is white, how would the error signal be chosen to enhance signals of period P? How long should the adaptive filter be in this application? Draw a block diagram of the resulting line enhancer.

(b) The settling time and mean-squared error are governed by the characteristics of the observations' covariance matrix. In this problem, however, we should not assume that the signal component is random. As $\mathbf{X} = \mathbf{s} + \mathbf{N}$, what is $\mathbf{K}_X$ when the noise has zero mean? Determine the smallest and largest eigenvalues of this matrix. Express the smallest settling time in terms of the signal-to-noise ratio.

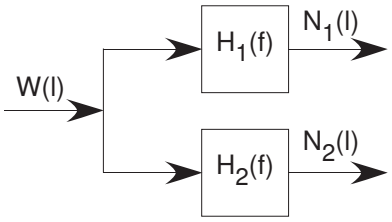
4.41 Adaptive Noise Cancellation

In adaptive noise cancellation, you use the *LMS* algorithm to estimate the noise rather than the signal. Once estimated, the noise can be subtracted from the observations, leaving behind the signal and the prediction residual.

$$X_1(l) = N_1(l)$$
$$X_2(l) = s(l) + N_2(l)$$

Here, the noise signals $N_1(l)$ and $N_2(l)$ are correlated Gaussian random processes. The idea is to use $X_1(l)$ to estimate $N_2(l)$ and then subtract this estimate from $X_2(l)$.

(a) The correlated noise signals can be modeled as a common white noise source passing through two filters in parallel.



Find the optimal non-causal Wiener filter's transfer function for adaptive noise cancellation. What is the resulting mean-square error?

(b) Develop an *LMS* filter for this problem and show the block diagram showing the signal flow.

(c) Simulate adaptive noise cancellation for the following scenario.

$$s(l) = \sin 2\pi f_0 l \ , f_0 = 0.25$$
$$H_1(f) = \frac{1}{1 - a_1 e^{-j2\pi f}} \ , a_1 = -0.8$$
$$H_2(f) = \frac{1}{1 - a_2 e^{-j2\pi f}} \ , a_2 = 0.6$$
$$\sigma_W^2 = 1$$

What range of values of μ work the best? Show the noisy signal and the result of the noise canceller. How does the mean-squared error compare with that of the optimal Wiener filter?

4.42 Optimal Spectral Estimation

While many spectral estimation procedures are found in the literature, few take into account the presence of additive noise. Assume that the observations consist of a signal s and statistically independent, additive, zero-mean noise N.

$$X(l) = s(l) + N(l), \ l = 0, \ldots, L-1$$

Treat finding the optimal estimate of the signal's spectrum as an optimal *FIR* filtering problem, where the quantity to be estimated is $\sum_l s(l) e^{-j2\pi f l}$.

(a) Find the spectral estimate that minimizes the mean-squared estimation error.

(b) Find this estimate's mean-squared error.

(c) Under what conditions is this estimate unbiased?

4.43 The covariance function estimate described on pages 114ff was biased (so the author claims) because the number of terms used at each lag varied without a corresponding variation in the normalization. Let's explore that claim closely. Assume that we now estimate the covariance function according to

$$\widehat{K}_X(m) = \frac{1}{D-|m|} \sum_{n=0}^{D-|m|-1} X(n)X(n+m), \quad 0 \le |m| \le D-1$$

(a) Find the expected value of this revised estimator, and show that it is indeed unbiased.

(b) To derive the variance of this estimate, we need the fourth moment of the observations, which is conveniently given in Chapter 2 (§2.1.9 {10}). Derive the covariance estimate's variance and determine whether it is consistent or not.

(c) Evaluate the expected value and variance of the spectral estimate corresponding to this covariance estimate.

(d) Does spectral estimate consistency become a reality with this new estimation procedure?

4.44 Let's see how spectral estimators work on a "real" dataset. The file `spectral.mat` contains a signal comprised of a sinusoid and additive noise.

(a) Evaluate the periodogram for this signal. What is your best guess for the sine wave's frequency and amplitude based on the periodogram? Are your estimates "good" in any sense?

(b) Use the Barlett spectral estimation procedure instead of the periodogram. Use a section length of 500, a Hanning window, and half-section overlap. Now what are your estimates of the sinusoid's frequency and amplitude?

(c) The default transform size in MATLAB's `fft` function is the data's length. Instead of using the section length in your Bartlett estimate, use a longer transform to determine the frequency resolution of the spectral peak presumably corresponding to the sinusoid. Compare the resolutions that section lengths of 500 and 1000 provide. Also compare the resolutions the Hanning and rectangular windows provide for these section lengths.

4.45 The data contained in `welldata.mat` is an actual recording of noise produced by a mechanical system. The variable `Fs` contains the sampling frequency. As with any data about which you know little, the transform length you ultimately use depends on what resolution you think you need.

(a) As a preliminary, produce a Gaussian process having a first-order spectrum (pole at 0.9; length = 2^{17}). Compute the periodogram for increasingly larger data lengths of 128, 1024, 8192, and 65,536. Does the spectrum become smoother as the length increases?

(b) Use the Bartlett spectral estimation procedure, with these lengths defining both the section and transform lengths. Now how does the error behave? Analytically predict how large the error should be. Do your simulations agree with the predictions?

(c) Now apply the Bartlett spectral estimate to the dataset. Find a section length that seems best.

(d) Compute the data's minimum variance spectral estimate using a section length of 128. Use a transform length of 1024. How does the minimum variance estimate compare with the Bartlett spectral estimate? Consider resolution and the apparent error.

4.46 Rather than the covariance matrix estimate given in Eq. 4.47 {122}, we can use a Bartlett-procedure-like maximum likelihood estimate. Imagine the length L observations segmented into K frames each having duration D. The k^{th} frame is formed into the column matrix $\mathbf{X}_k$ as $\text{col}[X(kD), X(kD+1), \ldots, X(kD+D-1)]$, $k = 0, \ldots, K-1$.

(a) Assuming the frames are statistically independent and are each distributed as a zero-mean Gaussian random vector having covariance matrix $\mathbf{K}_X$, find the log-likelihood function.

(b) By evaluating the gradient of this expression with respect to the matrix $\mathbf{K}$, find the maximum likelihood estimate of the covariance matrix $\widehat{\mathbf{K}}_{\mathrm{ML}}^{X}$.

(c) How does this estimate differ from that provided in Eq. 4.47 {122}?

Note: The matrix gradient identities provided in §B.3 {240} should prove useful.

4.47 The rather interesting form of the minimum variance spectral estimate (Eq. 4.47 {122}) has a periodogram counterpart.

(a) Show that the quadratic form $\mathbf{e}'(f)\widehat{\mathbf{K}}_X\mathbf{e}(f)$ equals the periodogram.

(b) Show that the Bartlett spectral estimate can be written as $\mathbf{e}'(f)\widehat{\mathbf{K}}_{\mathrm{ML}}^{X}\mathbf{e}(f)$, where $\widehat{\mathbf{K}}_{\mathrm{ML}}^{X}$ is the subject of Problem 4.46.

4.48 Accurately estimating the parameters of an *AR* model using linear prediction is contingent on zero-mean observations: We must use covariance rather than correlation matrix estimates to calculate these parameters. To illustrate this point, consider a sequence of observations $X(l)$ having mean m. Rather than subtracting the mean, you calculate the parameters presuming zero-mean data.

(a) How is the "covariance matrix" calculated by ignoring the mean related to the actual one? Relate the inverse of this matrix to the inverse of the true covariance matrix.

(b) Find an expression relating the true *AR* parameters to the computed ones. Show that the quantity $1/m^2 + \mathbf{1}^t\mathbf{K}^{-1}\mathbf{1}$ controls the deviation of the calculated parameters from the true ones.

4.49 Because of the special structure of the equation $\mathbf{K}_X\mathbf{a} = \sigma_w^2\boldsymbol{\delta}$ that defines the linear prediction parameters, the *AR* power spectrum can be directly related to the minimum variance power spectrum [13];[71: 354–56].

(a) Let $\mathbf{a}_r = \mathrm{col}[1, -a_1^r, \ldots, -a_r^r], r = 0, \ldots, p$, denote the solution to the r^{th}-order linear prediction problem. Form the matrix $\mathbf{A}$ having these vectors as its columns.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -a_1^p & 1 & \cdots & 0 & 0 \\ -a_2^p & -a_1^{p-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 1 & 0 \\ -a_p^p & -a_{p-1}^{p-1} & \cdots & -a_1^1 & 1 \end{bmatrix}$$

Show that $\mathbf{K}_X\mathbf{A} = \mathbf{U}$, where $\mathbf{U}$ is an upper triangular matrix. Here, $\mathbf{K}_X$ is a $(p+1)\times(p+1)$ matrix.

(b) Show that $\mathbf{A}^t\mathbf{K}_X\mathbf{A} = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix. What are its diagonal elements?

(c) Show that $\mathbf{K}_X^{-1} = \mathbf{A}\mathbf{D}^{-1}\mathbf{A}^t$.

(d) Let's put these matrix properties of linear prediction to use. Using the result of part (c) and the formula $\mathbf{e}'(f)\mathbf{K}_X^{-1}\mathbf{e}(f)$ for the reciprocal of the minimum variance spectrum, relate the spectra produced by the two techniques.

(e) From this relationship, predict which spectral estimation technique would have greater resolution.

4.50 The data contained in `spectral.mat` consist of a sinusoid contained in additive first-order, lowpass Gaussian noise. It would seem that the autoregressive spectral estimate would work well on such data.

(a) How many poles and how many zeros should it take to model the data? Does the autoregressive model indeed describe the data?

(b) Calculate autoregressive spectra for the dataset using model orders ranging from 1 to 50. Use the MATLAB function `xcorr` to calculate the correlation function estimate (use the `biased` option) and the function `levinson` to calculate the model parameters. Show the spectra in a `mesh` plot.

(c) Use the *AIC* and *MDL* methods to estimate the best model order. Do these model order estimates match that needed to discern the presence of the sinusoid in the spectra?

4.51 To illustrate the importance of *ARMA* models, consider a simple, but ubiquitous variant of the *AR* signal modeling [72]. We observe a *AR* signal $s(l)$ (order p) in the presence of statistically independent white noise $N(l)$ having variance σ_N^2.

(a) Express the power spectrum of the observations in terms of the signal's linear predictive parameters. Use this result to derive an equivalent *ARMA* model for the observations $X(l)$. What are the orders of this model?

The zeros introduced by the additive noise are undistinguished: Their locations depend entirely on the pole locations and on the noise variance σ_N^2. We seek an algorithm that might calculate the *AR* parameters despite the presence of additive noise.

(b) Show that for *any* $ARMA(p,q)$ model, the equations governing the covariance function of an *ARMA* process {127} obey $K_X(m) = \sum_{k=1}^{p} a_k K_X(m-k)$ for $m > q$, which are known as the *higher-order Yule-Walker equations*.*

(c) Based on this equation, develop an algorithm for determining the *AR* parameters of an *ARMA* process.

(d) How might you estimate the variance of the additive white noise?

4.52 Filter Coefficient Estimation

White Gaussian noise $W(l)$ serves as the input to a simple digital filter governed by the difference equation

$$X_l = aX_{l-1} + W_l \; .$$

We want to estimate the filter's coefficient a by processing the output observed over $l = 0, \ldots, L-1$. Prior to $l = 0$, the filter's input is zero.

(a) Find an estimate of a.

(b) What is the Cramér-Rao bound for your estimate?

4.53 The histogram probability density estimator is a special case of a more general class of estimators known as *kernel estimators*.

$$\hat{p}_X(x) = \frac{1}{L} \sum_{l=0}^{L-1} k\big(x - X(l)\big)$$

Here, the kernel $k(\cdot)$ is usually taken to be a density itself.

(a) What is the kernel for the histogram estimator?

(b) Interpret the kernel estimator in signal processing terminology. Predict what the most time consuming computation of this estimate might be. Why?

(c) Show that the sample average equals the expected value of a random variable having the density $\hat{p}_X(x)$ *regardless* of the choice of kernel.

4.54 Experiments in Compressive Sensing

In the compressive sensing (CS) framework, an unobserved signal $\mathbf{x}$, $L = \dim(\mathbf{x})$, has a K-sparse basis expansion: $\mathbf{x} = \mathbf{\Phi a}$. The columns of $\mathbf{\Phi}$ form an orthonormal basis and $\mathbf{a}$ are the expansion coefficients. To be K-sparse, the coefficient vector $\mathbf{a}$ has only K non-zero entries at unknown locations. M inner-product-type measurements are made with the $M \times L$ measurement matrix $\mathbf{\Psi}$: $\mathbf{y} = \mathbf{\Psi x}$. From these $M \ll L$ measurements, we want to construct $\mathbf{x}$.

According to CS theory, the number of measurements needed to construct $\mathbf{x}$ with high probability is of the order $O(K \log(L/K))$ when a "random" measurement matrix is used. The coefficient vector $\mathbf{a}$ is found as a solution of the ℓ_1 optimization problem

$$\min \|\mathbf{a}\|_1 \quad \text{subject to} \quad \mathbf{\Psi\Phi a} = \mathbf{y} \; .$$

In practice, we may not know how sparse the signal actually is. In this problem, we try to find how well solving the CS problem works in a scenario where the sparsity is known.

*The estimates calculated this way often have much larger variances than those computed from the usual Yule-Walker equations [82].

Software. The following Matlab commands produce a sparse length-N coefficient vector,[†] with random non-zero values and locations.

```
a = zeros(N, 1);
q = randperm(N);
a(q(1:K)) = randn(K,1);
```

We form a measurement matrix $\mathbf{\Psi}$ from Gaussian random numbers. NOTE: The *rows* of $\mathbf{\Psi}$ should form an orthonormal set. Use the Matlab function `orth()` to make the matrix have this property.

```
Psi = randn(M, L);
Psi = orth(Psi')';
```

The function `l1eq()` solves the constrained optimization problem and has the calling sequence

```
A = Psi*Phi;
% a0 = initial starting point, here the least-squares solution
a0 = A'*inv(A*A')*y;
a_hat = l1eq(a0,A,[],y);
```

(a) For the first experiment, let the signal itself be a sparse waveform, having $K = 20$ non-zero values. In this case, $N = L$; use $L = 512$. In this case, the matrix $\mathbf{\Phi}$ is the identity matrix. Using $M = 120$ measurements, compare the least-squares solution to the actual sparse signal for one example. Is the least-squares solution sparse?

(b) Solve the ℓ_1 optimization problem for the same data and measurements. Now is the solution sparse? Find the normalized root-mean-squared (rms) error (normalized by the signal's rms value) using the Matlab function `norm()`: ε^2 =`norm(a_hat-a)/norm(a)` and compare with the normalized rms error resulting from the least-squares solution.

(c) For the second round of experiments, let the signal be sparse with respect to a cosine-only Fourier series.

$$x_l = \sum_{j=0}^{L/2} a_j \sqrt{\frac{2}{L}} \cos \frac{2\pi jl}{L}, \quad l = 0, \ldots, L-1$$

Set up the matrix $\mathbf{\Phi}$ and solve the compressive sensing problem for the same values of K, L and M.

(d) The major issue is how many measurements are needed. Using the same sparse signal $\mathbf{x}$, find the normalized rms error as the number of measurements M varies from K to 120. Plot the normalized rms error as a function of M. Did you find anything surprising about your plot?

(e) Now set $K = 10$ and create a new signal that is sparse with respect to the Fourier basis. Plot the normalized error as in part (**d**). Does the threshold number of measurements necessary for accurate reconstruction follow the predicted behavior?

[†] In some cases, the signal's dimension (N) is smaller than its length.

Chapter 5

Detection Theory

5.1 Elementary Hypothesis Testing

In statistics, hypothesis testing is some times known as decision theory or simply testing. Here, one of several *models* $\mathscr{M}$ are presumed to describe a set of observed data, and you want to find which model best describes the observations. The key result around which all decision theory revolves is the likelihood ratio test.

5.1.1 The Likelihood Ratio Test

In a binary hypothesis testing problem, four possible outcomes can result. Model $\mathscr{M}_0$ did in fact represent the best model for the data and the decision rule said it was (a correct decision) or said it wasn't (an erroneous decision). The other two outcomes arise when model $\mathscr{M}_1$ was in fact true with either a correct or incorrect decision made. The decision process operates by segmenting the range of observation values into two disjoint *decision regions* $\Re_0$ and $\Re_1$. All values of $\mathbf{X}$ fall into either $\Re_0$ or $\Re_1$. If a given $\mathbf{X}$ lies in $\Re_0$, for example, we will announce our decision "model $\mathscr{M}_0$ was true"; if in $\Re_1$, model $\mathscr{M}_1$ would be proclaimed. To derive a rational method of deciding which model best describes the observations, we need a criterion to assess the quality of the decision process. Optimizing this criterion will specify the decision regions.

The *Bayes' decision criterion* seeks to minimize a cost function associated with making a decision. Let C_{ij} be the cost of mistaking model j for model i ($i \neq j$) and C_{ii} the presumably smaller cost of correctly choosing model i: $C_{ij} > C_{ii}, i \neq j$. Let π_i be the *a priori* probability of model i. The so-called *Bayes' cost* $\overline{C}$ is the average cost of making a decision.

$$\begin{aligned}\overline{C} &= \sum_{i,j} C_{ij} \Pr[\text{say } \mathscr{M}_i \text{ when } \mathscr{M}_j \text{ true}] \\ &= \sum_{i,j} C_{ij}\pi_j \Pr[\text{say } \mathscr{M}_i | \mathscr{M}_j \text{ true}]\end{aligned}$$

The Bayes' cost can be expressed as

$$\begin{aligned}\overline{C} &= \sum_{i,j} C_{ij}\pi_j \Pr[\mathbf{X} \in \Re_i | \mathscr{M}_j \text{ true}] = \sum_{i,j} C_{ij}\pi_j \int_{\Re_i} p_{\mathbf{X}|\mathscr{M}_j}(\mathbf{X}|\mathscr{M}_j)\, d\mathbf{X} \\ &= \int_{\Re_0} \{C_{00}\pi_0 p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0) + C_{01}\pi_1 p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)\}\, d\mathbf{X} \\ &\quad + \int_{\Re_1} \{C_{10}\pi_0 p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0) + C_{11}\pi_1 p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)\}\, d\mathbf{X}\end{aligned}$$

$p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{x}|\mathscr{M}_i)$ is the conditional probability density function of the observed data $\mathbf{X}$ given that model $\mathscr{M}_i$ was true. To minimize this expression with respect to the decision regions $\Re_0$ and $\Re_1$, ponder which integral would yield the smallest value if its integration domain included a specific observation vector. This selection

process defines the decision regions; for example, we choose $\mathscr{M}_0$ for those values of $\mathbf{X}$ which yield a smaller value for the first integral.

$$\Re_0 = \{\mathbf{X}\colon \pi_0 C_{00} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0) + \pi_1 C_{01} p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1) < \pi_0 C_{10} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0) + \pi_1 C_{11} p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)\}$$

We choose $\mathscr{M}_1$ when the inequality is reversed. This expression is easily manipulated to obtain the decision rule known as the *likelihood ratio test*.

$$\boxed{\frac{p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)}{p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\pi_0(C_{10}-C_{00})}{\pi_1(C_{01}-C_{11})}} \tag{5.1}$$

The comparison relation means selecting model $\mathscr{M}_1$ if the left-hand ratio exceeds the value on the right; otherwise, $\mathscr{M}_0$ is selected. Thus, the *likelihood ratio* $p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)/p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)$, symbolically represented by $\Lambda(\mathbf{X})$, is computed from the observed value of $\mathbf{X}$ and then compared with a *threshold* η equaling $\big[\pi_0(C_{10}-C_{00})\big]\big/\big[\pi_1(C_{01}-C_{11})\big]$. Thus, when two models are hypothesized, the likelihood ratio test can be succinctly expressed as the comparison of the likelihood ratio with a threshold.

$$\boxed{\Lambda(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \eta} \tag{5.2}$$

The data processing operations are captured entirely by the likelihood ratio $p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)/p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)$. Furthermore, note that only the value of the likelihood ratio *relative* to the threshold matters; to simplify the computation of the likelihood ratio, we can perform *any* positively monotonic operations simultaneously on the likelihood ratio and the threshold without affecting the comparison. We can multiply the ratio by a positive constant, add any constant, or apply a monotonically increasing function which simplifies the expressions. We single one such function, the logarithm, because it simplifies likelihood ratios that commonly occur in signal processing applications. Known as the log-likelihood, we explicitly express the likelihood ratio test with it as

$$\boxed{\ln\Lambda(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta}. \tag{5.3}$$

Useful simplifying transformations are problem-dependent; by laying bare that aspect of the observations essential to the model testing problem, we reveal the *sufficient statistic* $\Upsilon(\mathbf{X})$: the scalar quantity which best summarizes the data [65: pp. 18-22]. The likelihood ratio test is best expressed in terms of the sufficient statistic.

$$\boxed{\Upsilon(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma} \tag{5.4}$$

We will denote the threshold value by γ when the sufficient statistic is used or by η when the likelihood ratio appears prior to its reduction to a sufficient statistic.

As we shall see, if we use a different criterion other than the Bayes' criterion, the decision rule often involves the likelihood ratio. The likelihood ratio is comprised of the quantities $p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{x}|\mathscr{M}_i)$, termed the *likelihood function*, which is also important in estimation theory. It is this conditional density that portrays the probabilistic model describing data generation. The likelihood function completely characterizes the kind of "world" assumed by each model; for each model, we must specify the likelihood function so that we can solve the hypothesis testing problem.

A complication, which arises in some cases, is that the sufficient statistic may not be monotonic. If monotonic, the decision regions $\Re_0$ and $\Re_1$ are simply connected (all portions of a region can be reached without crossing into the other region). If not, the regions are not simply connected and decision region islands are created (see Problem 5.2). Such regions usually complicate calculations of decision performance. Monotonic or not, the decision rule proceeds as described: the sufficient statistic is computed for each observation vector and compared to a threshold.

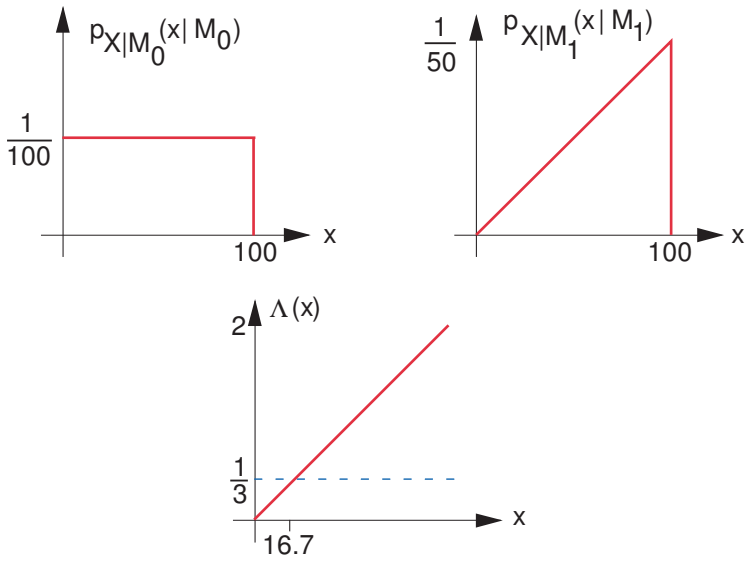


Figure 5.1: Conditional densities for the grade distributions assuming that a student did not study ($\mathscr{M}_0$) or did ($\mathscr{M}_1$) are shown in the top row. The lower portion depicts the likelihood ratio formed from these densities.

Example

An instructor in a course in detection theory wants to determine if a particular student studied for his last test. The observed quantity is the student's grade, which we denote by X. Failure may not indicate studiousness: conscientious students may fail the test. Define the models as

$$\begin{aligned} \mathscr{M}_0&: \quad \text{did not study} \\ \mathscr{M}_1&: \quad \text{studied} \end{aligned}$$

The conditional densities of the grade are shown in Fig. 5.1. Based on knowledge of student behavior, the instructor assigns *a priori* probabilities of $\pi_0 = 1/4$ and $\pi_1 = 3/4$. The costs C_{ij} are chosen to reflect the instructor's sensitivity to student feelings: $C_{01} = 1 = C_{10}$ (an erroneous decision either way is given the same cost) and $C_{00} = 0 = C_{11}$. The likelihood ratio is plotted in Fig. 5.1 and the threshold value η, which is computed from the *a priori* probabilities and the costs to be $1/3$, is indicated. The calculations of this comparison can be simplified in an obvious way.

$$\frac{X}{50} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{1}{3} \quad \text{or} \quad X \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{50}{3} = 16.7$$

The multiplication by the factor of 50 is a simple illustration of the reduction of the likelihood ratio to a sufficient statistic. Based on the assigned costs and *a priori* probabilities, the optimum decision rule says the instructor must assume that the student did not study if the student's grade is less than 16.7; if greater, the student is assumed to have studied despite receiving an abysmally low grade such as 20. Note that as the densities given by each model overlap entirely: the possibility of making the wrong interpretation *always* haunts the instructor. However, no other procedure will be better!

5.1.2 Criteria in Hypothesis Testing

The criterion used in the previous section—minimize the average cost of an incorrect decision—may seem to be a contrived way of quantifying decisions. Well, often it is. For example, the Bayesian decision rule depends

explicitly on the *a priori* probabilities; a rational method of assigning values to these—either by experiment or through true knowledge of the relative likelihood of each model—may be unreasonable. In this section, we develop alternative decision rules that try to answer such objections. One essential point will emerge from these considerations: *the fundamental nature of the decision rule does not change with choice of optimization criterion*. Even criteria remote from error measures can result in the likelihood ratio test (see Problem 5.4). Such results do not occur often in signal processing and underline the likelihood ratio test's significance.

Maximum Probability of a Correct Decision

As only one model can describe any given set of data (the models are mutually exclusive), the probability of being correct P_c for distinguishing two models is given by

$$P_c = \Pr[\text{say } \mathcal{M}_0 \text{ when } \mathcal{M}_0 \text{ true}] + \Pr[\text{say } \mathcal{M}_1 \text{ when } \mathcal{M}_1 \text{ true}].$$

We wish to determine the optimum decision region placement by maximizing P_c. Expressing the probability correct in terms of the likelihood functions $p_{\mathbf{X}|\mathcal{M}_i}(\mathbf{x}|\mathcal{M}_i)$, the *a priori* probabilities, and the decision regions,

$$P_c = \int_{\Re_0} \pi_0 p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{x}|\mathcal{M}_0)\, d\mathbf{X} + \int_{\Re_1} \pi_1 p_{\mathbf{x}|\mathcal{M}_1}(\mathbf{x}|\mathcal{M}_1)\, d\mathbf{X}.$$

We want to maximize P_c by selecting the decision regions $\Re_0$ and $\Re_1$. The probability correct is maximized by associating each value of $\mathbf{X}$ with the largest term in the expression for P_c. Decision region $\Re_0$, for example, is defined by the collection of values of $\mathbf{X}$ for which the first term is largest. As all of the quantities involved are non-negative, the decision rule maximizing the probability of a correct decision is

> Given $\mathbf{X}$, choose $\mathcal{M}_i$ for which the product $\pi_i p_{\mathbf{X}|\mathcal{M}_i}(\mathbf{x}|\mathcal{M}_i)$ is largest.

Simple manipulations lead to the likelihood ratio test.

$$\frac{p_{\mathbf{x}|\mathcal{M}_1}(\mathbf{x}|\mathcal{M}_1)}{p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{x}|\mathcal{M}_0)} \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} \frac{\pi_0}{\pi_1}$$

Note that if the Bayes' costs were chosen so that $C_{ii} = 0$ and $C_{ij} = C$, $(i \neq j)$, we would have the same threshold as in the previous section.

To evaluate the quality of the decision rule, we usually compute the *probability of error* P_e rather than the probability of being correct. This quantity can be expressed in terms of the observations, the likelihood ratio, and the sufficient statistic.

$$\begin{aligned} P_e &= \pi_0 \int_{\Re_1} p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{x}|\mathcal{M}_0)\, d\mathbf{X} + \pi_1 \int_{\Re_0} p_{\mathbf{x}|\mathcal{M}_1}(\mathbf{x}|\mathcal{M}_1)\, d\mathbf{X} \\ &= \pi_0 \int_{\Lambda>\eta} p_{\Lambda|\mathcal{M}_0}(\Lambda|\mathcal{M}_0)\, d\Lambda + \pi_1 \int_{\Lambda<\eta} p_{\Lambda|\mathcal{M}_1}(\Lambda|\mathcal{M}_1)\, d\Lambda \\ &= \pi_0 \int_{\Upsilon>\gamma} p_{\Upsilon|\mathcal{M}_0}(\Upsilon|\mathcal{M}_0)\, d\Upsilon + \pi_1 \int_{\Upsilon<\gamma} p_{\Upsilon|\mathcal{M}_1}(\Upsilon|\mathcal{M}_1)\, d\Upsilon \end{aligned} \tag{5.5}$$

When the likelihood ratio is non-monotonic, the first expression is most difficult to evaluate. When monotonic, the middle expression proves the most difficult. Furthermore, these expressions point out that the likelihood ratio and the sufficient statistic can be considered a function of the observations $\mathbf{X}$; hence, they are random variables and have probability densities for each model. Another aspect of the resulting probability of error is that *no other decision rule can yield a lower probability of error*. This statement is obvious as we minimized the probability of error in deriving the likelihood ratio test. The point is that these expressions represent a lower bound on performance (as assessed by the probability of error). This probability will be non-zero if the conditional densities overlap over some range of values of $\mathbf{X}$, such as occurred in the previous example. In this region of overlap, the observed values are ambiguous: either model is consistent with the observations. Our "optimum" decision rule operates in such regions by selecting that model which is most likely (has the highest probability) of generating any particular value.

Neyman-Pearson Criterion

Situations occur frequently where assigning or measuring the *a priori* probabilities P_i is unreasonable. For example, just what is the *a priori* probability of a supernova occurring in any particular region of the sky? We clearly need a model evaluation procedure which can function without *a priori* probabilities. This kind of test results when the so-called Neyman-Pearson criterion is used to derive the decision rule. The ideas behind and decision rules derived with the Neyman-Pearson criterion [76] will serve us well in sequel; their result is important!

Using nomenclature from radar, where model $\mathcal{M}_1$ represents the presence of a target and $\mathcal{M}_0$ its absence, the various types of correct and incorrect decisions have the following names [98: pp. 127–9].*

$P_D = \Pr[\text{say } \mathcal{M}_1 \vert \mathcal{M}_1 \text{ true}]$	*Detection*–we say it's there when it is
$P_F = \Pr[\text{say } \mathcal{M}_1 \vert \mathcal{M}_0 \text{ true}]$	*False-alarm*–we say it's there when it's not
$P_M = \Pr[\text{say } \mathcal{M}_0 \vert \mathcal{M}_1 \text{ true}]$	*Miss*–we say it's not there when it is

The remaining probability $\Pr[\text{say } \mathcal{M}_0 \vert \mathcal{M}_0 \text{ true}]$ has historically been left nameless and equals $1 - P_F$. We should also note that the detection and miss probabilities are related by $P_M = 1 - P_D$. As these are conditional probabilities, they do not depend on the *a priori* probabilities and the two probabilities P_F and P_D characterize the errors when *any* decision rule is used.

These two probabilities are related to each other in an interesting way. Expressing these quantities in terms of the decision regions and the likelihood functions, we have

$$P_F = \int_{\Re_1} p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{x}|\mathcal{M}_0)\, d\mathbf{X}, \qquad P_D = \int_{\Re_1} p_{\mathbf{x}|\mathcal{M}_1}(\mathbf{x}|\mathcal{M}_1)\, d\mathbf{X}\,.$$

As the region $\Re_1$ shrinks, *both* of these probabilities tend toward zero; as $\Re_1$ expands to engulf the entire range of observation values, they both tend toward unity. This rather direct relationship between P_D and P_F does not mean that they equal each other; in most cases, as $\Re_1$ expands, P_D increases more rapidly than P_F (we had better be right more often than we are wrong!). However, the "ultimate" situation where a rule is always right and never wrong ($P_D = 1$, $P_F = 0$) cannot occur when the conditional distributions overlap. Thus, to increase the detection probability we must also allow the false-alarm probability to increase. This behavior represents the fundamental tradeoff in hypothesis testing and detection theory.

One can attempt to impose a performance criterion that depends only on these probabilities with the consequent decision rule not depending on the *a priori* probabilities. The Neyman-Pearson criterion assumes that the false-alarm probability is constrained to be less than or equal to a specified value α while we attempt to maximize the detection probability P_D.

$$\max_{\Re_1} P_D \;\text{ subject to } P_F \leq \alpha$$

A subtlety of the succeeding solution is that the underlying probability distribution functions may not be continuous, with the result that P_F can never equal the constraining value α. Furthermore, an (unlikely) possibility is that the optimum value for the false-alarm probability is somewhat less than the criterion value. Assume, therefore, that we rephrase the optimization problem by requiring that the false-alarm probability equal a value α' that is less than or equal to α.

This optimization problem can be solved using Lagrange multipliers (see §3.2.1 {51}); we seek to find the decision rule that maximizes

$$F = P_D + \lambda(P_F - \alpha'),$$

where λ is the Lagrange multiplier. This optimization technique amounts to finding the decision rule that maximizes F, then finding the value of the multiplier that allows the criterion to be satisfied. As is usual in the derivation of optimum decision rules, we maximize these quantities with respect to the decision regions.

*In hypothesis testing, a false-alarm is known as a type I error and a miss a type II error.

Expressing P_D and P_F in terms of them, we have

$$\begin{aligned} F &= \int_{\Re_1} p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)\,d\mathbf{X} + \lambda\left(\int_{\Re_1} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)d\mathbf{X} - \alpha'\right) \\ &= -\lambda\alpha' + \int_{\Re_1}\left[p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1) + \lambda p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)\right] d\mathbf{X}. \end{aligned}$$

To maximize this quantity with respect to $\Re_1$, we need only to integrate over those regions of $\mathbf{X}$ where the integrand is positive. The region $\Re_1$ thus corresponds to those values of $\mathbf{X}$ where $p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1) > -\lambda p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)$ and the resulting decision rule is

$$\frac{p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)}{p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} -\lambda$$

The ubiquitous likelihood ratio test again appears; it *is* indeed the fundamental quantity in hypothesis testing. Using the logarithm of the likelihood ratio or the sufficient statistic, this result can be expressed as either

$$\ln\Lambda(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln(-\lambda) \quad \text{or} \quad \Upsilon(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma.$$

We have not as yet found a value for the threshold. The false-alarm probability can be expressed in terms of the Neyman-Pearson threshold in two (useful) ways.

$$\boxed{\begin{aligned} P_F &= \int_{-\lambda}^{\infty} p_{\Lambda|\mathscr{M}_0}(\Lambda|\mathscr{M}_0)\,d\Lambda \\ &= \int_{\gamma}^{\infty} p_{\Upsilon|\mathscr{M}_0}(\Upsilon|\mathscr{M}_0)\,d\Upsilon \end{aligned}} \tag{5.6}$$

One of these implicit equations must be solved for the threshold by setting P_F equal to α'. The selection of which to use is usually based on pragmatic considerations: the easiest to compute. From the previous discussion of the relationship between the detection and false-alarm probabilities, we find that to maximize P_D we must allow α' to be as large as possible while remaining less than α. Thus, we want to find the *smallest* value of $-\lambda$ (note the minus sign) consistent with the constraint. Computation of the threshold is problem-dependent, but a solution always exists.

Example

An important application of the likelihood ratio test occurs when $\mathbf{X}$ is a Gaussian random vector for each model. Suppose the models correspond to Gaussian random vectors having different mean values but sharing the same identity covariance.

$$\begin{aligned} \mathscr{M}_0 &: \mathbf{X} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathscr{M}_1 &: \mathbf{X} \sim \mathscr{N}(\mathbf{m}, \sigma^2\mathbf{I}) \end{aligned}$$

Thus, $\mathbf{X}$ is of dimension L and has statistically independent, equal variance components. The vector of means $\mathbf{m} = \text{col}[m_0, \ldots, m_{L-1}]$ distinguishes the two models. The likelihood functions associated this problem are

$$\begin{aligned} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0) &= \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{X_l}{\sigma}\right)^2\right\} \\ p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1) &= \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{X_l - m_l}{\sigma}\right)^2\right\} \end{aligned}$$

The likelihood ratio $\Lambda(\mathbf{X})$ becomes

$$\Lambda(\mathbf{X}) = \frac{\prod_{l=0}^{L-1} \exp\left\{-\frac{1}{2}\left(\frac{X_l - m_l}{\sigma}\right)^2\right\}}{\prod_{l=0}^{L-1} \exp\left\{-\frac{1}{2}\left(\frac{X_l}{\sigma}\right)^2\right\}}$$

This expression for the likelihood ratio is complicated. In the Gaussian case (and many others), we use the logarithm the reduce the complexity of the likelihood ratio and form a sufficient statistic.

$$\begin{aligned}\ln\Lambda(\mathbf{X}) &= \sum_{l=0}^{L-1}\left\{-\frac{1}{2}\frac{(X_l - m_l)^2}{\sigma^2} + \frac{1}{2}\frac{X_l^2}{\sigma^2}\right\} \\ &= \frac{1}{\sigma^2}\sum_{l=0}^{L-1} m_l X_l - \frac{1}{2\sigma^2}\sum_{l=0}^{L-1} m_l^2\end{aligned}$$

The likelihood ratio test then has the much simpler, but equivalent form

$$\sum_{l=0}^{L-1} m_l X_l \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta + \frac{1}{2}\sum_{l=0}^{L-1} m_l^2 .$$

To focus on the model evaluation aspects of this problem, let's assume means be equal to a positive constant: $m_l = m\ (>0)$.*

$$\sum_{l=0}^{L-1} X_l \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\sigma^2}{m}\ln\eta + \frac{Lm}{2}$$

Note that all that need be known about the observations $\{X_l\}$ is their sum. This quantity is the sufficient statistic for the Gaussian problem: $\Upsilon(\mathbf{X}) = \sum X_l$ and $\gamma = \sigma^2 \ln\eta/m + Lm/2$.

When trying to compute the probability of error or the threshold in the Neyman-Pearson criterion, we must find the conditional probability density of one of the decision statistics: the likelihood ratio, the log-likelihood, or the sufficient statistic. The log-likelihood and the sufficient statistic are quite similar in this problem, but clearly we should use the latter. One practical property of the sufficient statistic is that it usually simplifies computations. For this Gaussian example, the sufficient statistic is a Gaussian random variable under each model.

$$\begin{aligned}\mathscr{M}_0&\colon \Upsilon(\mathbf{X}) \sim \mathscr{N}(0, L\sigma^2) \\ \mathscr{M}_1&\colon \Upsilon(\mathbf{X}) \sim \mathscr{N}(Lm, L\sigma^2)\end{aligned}$$

To find the probability of error from the expressions found on Page 152, we must evaluate the area under a Gaussian probability density function. These integrals are succinctly expressed in terms of $Q(x)$, which denotes the probability that a unit-variance, zero-mean Gaussian random variable exceeds x (see chapter 2 {8}). As $1 - Q(x) = Q(-x)$, the probability of error can be written as

$$P_e = \pi_1 Q\left(\frac{Lm - \gamma}{\sqrt{L}\sigma}\right) + \pi_0 Q\left(\frac{\gamma}{\sqrt{L}\sigma}\right).$$

An interesting special case occurs when $\pi_0 = 1/2 = \pi_1$. In this case, $\gamma = Lm/2$ and the probability of error becomes

$$P_e = Q\left(\frac{\sqrt{L}m}{2\sigma}\right).$$

As $Q(\cdot)$ is a monotonically decreasing function, the probability of error decreases with increasing values of the ratio $\sqrt{L}m/2\sigma$. However, as shown in appendix Fig. 2.1 {9}, $Q(\cdot)$ decreases in a nonlinear

*Why did the author assume that the mean was positive? What would happen if it were negative?

x	$Q^{-1}(x)$
10^{-1}	1.281
10^{-2}	2.396
10^{-3}	3.090
10^{-4}	3.719
10^{-5}	4.265
10^{-6}	4.754

Table 5.1: The table displays interesting values for $Q^{-1}(\cdot)$ that can be used to determine thresholds in the Neyman-Pearson variant of the likelihood ratio test. Note how little the inverse function changes for decade changes in its argument; $Q(\cdot)$ is indeed *very* nonlinear.

fashion. Thus, increasing m by a factor of two may decrease the probability of error by a larger *or* a smaller factor; the amount of change depends on the initial value of the ratio.

To find the threshold for the Neyman-Pearson test from the expressions given on Page 154, we need the area under a Gaussian density.

$$P_F = Q\left(\frac{\gamma}{\sqrt{L\sigma^2}}\right) = \alpha' \tag{5.7}$$

As $Q(\cdot)$ is a monotonic and continuous function, we can now set α' equal to the criterion value α with the result

$$\gamma = \sqrt{L}\sigma Q^{-1}(\alpha)\,.$$

where $Q^{-1}(\cdot)$ denotes the inverse function of $Q(\cdot)$. The solution of this equation cannot be performed analytically as no closed form expression exists for $Q(\cdot)$ (much less its inverse function); the criterion value must be found from tables or numerical routines. Because Gaussian problems arise frequently, the accompanying table provides numeric values for this quantity at the decade points. The detection probability is given by

$$P_D = Q\left(Q^{-1}(\alpha) - \frac{\sqrt{L}m}{\sigma}\right).$$

5.1.3 Performance Evaluation

We alluded earlier {153} to the relationship between the false-alarm probability P_F and the detection probability P_D as one varies the decision region. Because the Neyman-Pearson criterion depends on specifying the false-alarm probability to yield an acceptable detection probability, we need to examine carefully how the detection probability is affected by a specification of the false-alarm probability. The usual way these quantities are discussed is through a parametric plot of P_D versus P_F: the *receiver operating characteristic* or ROC.

As we discovered in the Gaussian example {154}, the sufficient statistic provides the simplest way of computing these probabilities; thus, they are usually considered to depend on the threshold parameter γ. In these terms, we have

$$P_D = \int_\gamma^\infty p_{\Upsilon|\mathscr{M}_1}(\Upsilon|\mathscr{M}_1)\,d\Upsilon \quad \text{and} \quad P_F = \int_\gamma^\infty p_{\Upsilon|\mathscr{M}_0}(\Upsilon|\mathscr{M}_0)\,d\Upsilon\,. \tag{5.8}$$

These densities and their relationship to the threshold γ are shown in Fig. 5.2. We see that the detection probability is greater than or equal to the false-alarm probability. Since these probabilities must decrease monotonically as the threshold is increased, the ROC curve must be concave-down and must *always* exceed the equality line (Fig. 5.3). The degree to which the ROC departs from the equality line $P_D = P_F$ measures the

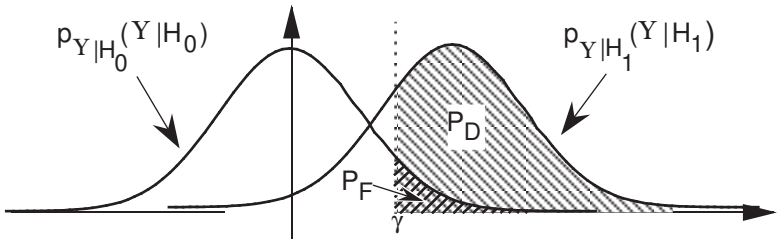


Figure 5.2: The densities of the sufficient statistic $\Upsilon(\mathbf{X})$ conditioned on two hypotheses are shown for the Gaussian example. The threshold γ used to distinguish between the two models is indicated. The false-alarm probability is the area under the density corresponding to $\mathscr{M}_0$ to the right of the threshold; the detection probability is the area under the density corresponding to $\mathscr{M}_1$.

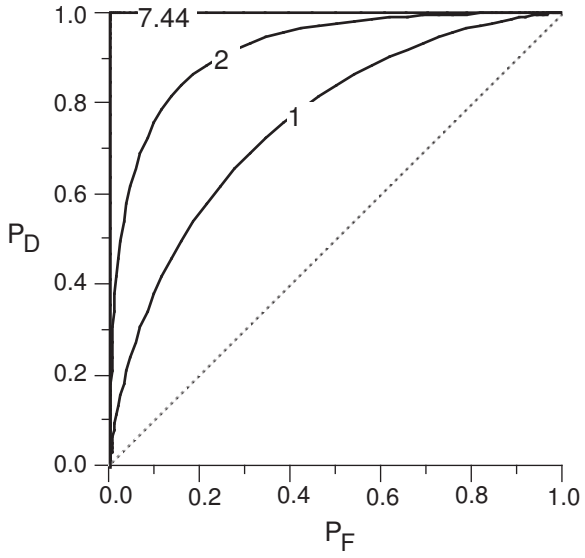


Figure 5.3: A plot of the receiver operating characteristic for the densities shown in the previous figure. Three ROC curves are shown corresponding to different values for the parameter $\sqrt{L}m/\sigma$.

relative "distinctiveness" between the two hypothesized models for generating the observations. In the limit, the two models can be distinguished perfectly if the ROC is discontinuous and consists of the point $(1,0)$. The two are totally confused if the ROC lies on the equality line (this would mean, of course, that the two models are identical); distinguishing the two in this case would be "somewhat difficult".

Example

Consider the Gaussian example we have been discussing where the two models differ only in the means of the conditional distributions. In this case, the two model-testing probabilities are given by

$$P_F = Q\left(\frac{\gamma}{\sqrt{L}\sigma}\right) \quad \text{and} \quad P_D = Q\left(\frac{\gamma - Lm}{\sqrt{L}\sigma}\right).$$

By re-expressing γ as $\frac{\sigma^2}{m}\gamma' + \frac{Lm}{2}$, we discover that these probabilities depend only on the ratio $\sqrt{L}m/\sigma$.

$$P_F = Q\left(\frac{\gamma'}{\sqrt{L}m/\sigma} + \frac{\sqrt{L}m}{2\sigma}\right), \qquad P_D = Q\left(\frac{\gamma'}{\sqrt{L}m/\sigma} - \frac{\sqrt{L}m}{2\sigma}\right)$$

As this signal-to-noise ratio increases, the ROC curve approaches its "ideal" form: the northwest corner of a square as illustrated in Fig. 5.3 by the value of 7.44 for $\sqrt{L}m/\sigma$, which corresponds to a signal-to-noise ratio of $7.44^2 \approx 17$ dB. If a small false-alarm probability (say 10^{-4}) is specified, a large detection probability (0.9999) can result. Such values of signal-to-noise ratios can thus be considered "large" and the corresponding model evaluation problem relatively easy. If, however, the signal-to-noise ratio equals 4 (6 dB), the figure illustrates the worsened performance: a 10^{-4} specification on the false-alarm probability would result in a detection probability of essentially zero. Thus, in a fairly small signal-to-noise ratio range, the likelihood ratio test's performance capabilities can vary dramatically. However, no other decision rule can yield better performance.

Specification of the false-alarm probability for a new problem requires experience. Choosing a "reasonable" value for the false-alarm probability in the Neyman-Pearson criterion depends strongly on the problem difficulty. Too small a number will result in small detection probabilities; too large and the detection probability will be close to unity, suggesting that fewer false alarms could have been tolerated. Problem difficulty is assessed by the degree to which the conditional densities $p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)$ and $p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)$ overlap, a problem dependent measurement. If we are testing whether a distribution has one of two possible mean values as in our Gaussian example, a quantity like a signal-to-noise ratio will probably emerge as determining performance. The performance in this case can vary drastically depending on whether the signal-to-noise ratio is large or small. In other kinds of problems, the best possible performance provided by the likelihood ratio test can be poor. For example, consider the problem of determining which of two zero-mean probability densities describes a given set of data consisting of statistically independent observations (Problem 5.2). Presumably, the variances of these two densities are equal as we are trying to determine which density is most appropriate. In this case, the performance probabilities can be quite low, especially when the general shapes of the densities are similar. Thus a single quantity, like the signal-to-noise ratio, does *not* emerge to characterize problem difficulty in all hypothesis testing problems. In sequel, we will analyze each model evaluation and detection problem in a standard way. After the sufficient statistic has been found, we will seek a value for the threshold that attains a specified false-alarm probability. The detection probability will then be determined as a function of "problem difficulty", the measure of which is problem-dependent. We can control the choice of false-alarm probability; we cannot control over problem difficulty. Confusingly, the detection probability will vary with *both* the specified false-alarm probability and the problem difficulty.

We are implicitly assuming that we have a rational method for choosing the false-alarm probability criterion value. In signal processing applications, we usually make a sequence of decisions and pass them to systems making more global determinations. For example, in digital communications problems the model evaluation formalism could be used to "receive" each bit. Each bit is received in sequence and then passed to the decoder which invokes error-correction algorithms. The important notions here are that the decision-making process occurs at a given *rate* and that the decisions are presented to other signal processing systems. The rate at which errors occur in system input(s) greatly influences system design. Thus, the selection of a false-alarm probability is usually governed by the *error rate* that can be tolerated by succeeding systems. If the decision rate is one per day, then a moderately large (say 0.1) false-alarm probability might be appropriate. If the decision rate is a million per second as in a one megabit communication channel, the false-alarm probability should be much lower: 10^{-12} would suffice for the one-tenth per day error rate.

5.1.4 Beyond Two Models

Frequently, more than two viable models for data generation can be defined for a given situation. The *classification* problem is to determine which of several models best "fits" a set of measurements. For example, determining the type of airplane from its radar returns forms a classification problem. The model evaluation framework has the right structure if we can allow more than two model. We happily note that in deriving the likelihood ratio test we did not need to assume that only two possible descriptions exist. Go back and examine the expression for the maximum probability correct decision rule $\{152\}$. If K models seem appropriate for a

specific problem, the decision rule maximizing the probability of making a correct choice is

$$\boxed{\text{Choose the largest of } \pi_i p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{x}|\mathscr{M}_i), \quad i=1,\ldots,K.}$$

To determine the largest of K quantities, exactly $K-1$ numeric comparisons need be made. When we have two possible models ($K=2$), this decision rule reduces to the computation of the likelihood ratio and its comparison to a threshold. In general, $K-1$ likelihood ratios need to be computed and compared to a threshold. Thus the likelihood ratio test can be viewed as a specific method for determining the largest of the decision statistics $\pi_i p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{x}|\mathscr{M}_i)$.

Since we need only the relative ordering of the K decision statistics to make a decision, we can apply any transformation $T(\cdot)$ to them that does not affect ordering. In general, possible transformations must be positively monotonic to satisfy this condition. For example, the needless common additive components in the decision statistics can be eliminated, even if they depend on the observations. Mathematically, "common" means that the quantity does not depend on the model index i. The transformation in this case would be of the form $T(z_i) = z_i - a$, clearly a monotonic transformation. A *positive* multiplicative factor can also the "canceled"; if negative, the ordering would be reversed and that cannot be allowed. The simplest resulting expression becomes the sufficient statistic $\Upsilon_i(\mathbf{X})$ for the model. Expressed in terms of the sufficient statistic, the maximum probability correct or the Bayesian decision rule becomes

$$\boxed{\text{Choose the largest of } C_i + \Upsilon_i(\mathbf{X}), \quad i=1,\ldots,K},$$

where C_i summarizes all additive terms that do not depend on the observation vector $\mathbf{X}$. The quantity $\Upsilon_i(\mathbf{X})$ is termed the *sufficient statistic associated with model i*. In many cases, the functional form of the sufficient statistic varies little from one model to another and expresses the necessary operations that summarize the observations. The constants C_i are usually lumped together to yield the threshold against which we compare the sufficient statistic. For example, in the binary model situation, the decision rule becomes

$$\Upsilon_1(\mathbf{X}) + C_1 \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \Upsilon_0(\mathbf{X}) + C_0 \quad \text{or} \quad \Upsilon_1(\mathbf{X}) - \Upsilon_0(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} C_0 - C_1.$$

Thus, the sufficient statistic for the decision rule is $\Upsilon_1(\mathbf{X}) - \Upsilon_0(\mathbf{X})$ and the threshold γ is $C_0 - C_1$.

Example

In the Gaussian problem just discussed, the logarithm of the likelihood function is

$$\ln p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{x}|\mathscr{M}_i) = -\frac{L}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{l=0}^{L-1}(X_l - m^{(i)})^2,$$

where $m^{(i)}$ is the mean under model i. After appropriate simplification that retains the ordering, we have

$$\Upsilon_i(\mathbf{X}) = \frac{m^{(i)}}{\sigma^2}\sum_{l=0}^{L-1} X_l \qquad C_i = -\frac{1}{2}\frac{Lm^{(i)2}}{\sigma^2} + c_i.$$

The term c_i is a constant defined by the error criterion; for the maximum probability correct criterion, this constant is $\ln \pi_i$.

When employing the Neyman-Pearson test, we need to specify the various error probabilities Pr[say $\mathscr{M}_i|\mathscr{M}_j$ true]. These specifications amount to determining the constants c_i when the sufficient statistic is used. Since $K-1$ comparisons will be used to home in on the optimal decision, only $K-1$ error probabilities need be specified. Typically, the quantities Pr[say $\mathscr{M}_i|\mathscr{M}_0$ true], $i=1,\ldots,K-1$, are used, particularly when the model $\mathscr{M}_0$ represents the situation when no signal is present (see Problem 5.7).

5.1.5 Model Consistency Testing

In many situations, we seek to check consistency of the observations with some preconceived model. Alternative models are usually difficult to describe parametrically since inconsistency may be beyond our modeling capabilities. We need a test that accepts consistency of observations with a model or rejects the model without pronouncing a more favored alternative. Assuming we know (or presume to know) the probability distribution of the observations under $\mathscr{M}_0$, the models are

$$\mathscr{M}_0 : \mathbf{X} \sim p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)$$
$$\mathscr{M}_1 : \mathbf{X} \not\sim p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)$$

Null hypothesis testing seeks to determine if the observations are consistent with this description. The best procedure for consistency testing amounts to determining whether the observations lie in a highly probable region as defined by the null probability distribution. However, no one region defines a probability that is less than unity. We must restrict the size of the region so that it best represents those observations maximally consistent with the model while satisfying a performance criterion. Letting P_F be a false-alarm probability established by us, we define the decision region $\Re_0$ to satisfy

$$\Pr[\mathbf{X} \in \Re_0|\mathscr{M}_0] = \int_{\Re_0} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)\, d\mathbf{X} = 1 - P_F \quad \text{and} \quad \min_{\Re_0} \int_{\Re_0} d\mathbf{X}.$$

Usually, this region is located about the mean, but may not be symmetrically centered if the probability density is skewed. Our null hypothesis test for model consistency becomes

$$\boxed{\begin{array}{l} \mathbf{X} \in \Re_0 \Longrightarrow \text{“say observations are consistent”} \\ \mathbf{X} \notin \Re_0 \Longrightarrow \text{“say observations are not consistent”} \end{array}}.$$

Example

Consider the problem of determining whether the sequence X_l, $l = 1, \ldots, L$, is white and Gaussian with zero mean and unit variance. Stated this way, the alternative model is not provided: is this model correct or not? We could estimate the probability density function of the observations and test the estimate for consistency. Here we take the null-hypothesis testing approach of converting this problem into a one-dimensional one by considering the statistic $X = \sum_l X_l^2$, which has a χ^2_L distribution. Because this probability distribution is unimodal, the decision region can be safely assumed to be an interval $[X', X'']$.* In this case, we can find an analytic solution to the problem of determining the decision region. Letting $X = X'' - X'$ denote the width of the interval, we seek the solution of the constrained optimization problem

$$\min_{X'} X \quad \text{subject to} \quad P_X(X' + X) - P_X(X') = 1 - P_F\,.$$

We convert the constrained problem into an unconstrained one using Lagrange multipliers.

$$\min_{X'} \left\{ X + \lambda \left[P_X(X' + X) - P_X(X') - (1 - P_F) \right] \right\}$$

Evaluation of the derivative of this quantity with respect to X' yields the result $p_X(X' + X) = p_X(X')$: to minimize the interval's width, the probability density function's values at the interval's endpoints must be equal. Finding these endpoints to satisfy the constraints amounts to searching the probability distribution at such points for increasing values of X until the required probability is contained within. For $L = 100$ and $P_F = 0.05$, the optimal decision region for the χ^2_L distribution is $[78.82, 128.5]$. Fig. 5.4 demonstrates ten testing trials for observations that fit the model and for observations that don't.

---

*This one-dimensional result for the consistency test may extend to the multi-dimensional case in the obvious way.

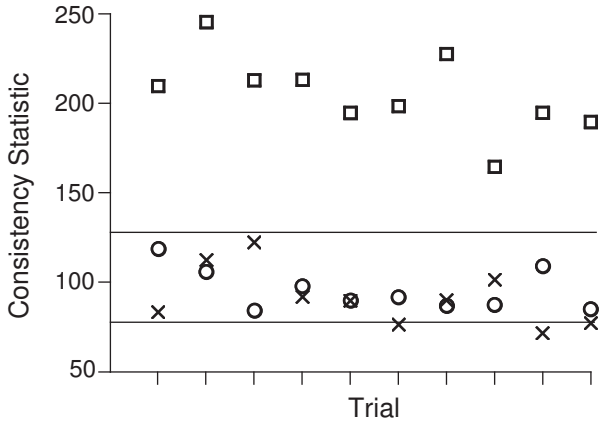


Figure 5.4: Ten trials of testing a 100-element sequence for consistency with a white, Gaussian model—$X_l \sim \mathcal{N}(0,1)$—for three situations. In the first (shown by the circles), the observations do conform to the model. In the second (boxes), the observations are zero-mean Gaussian but with variance two. Finally, the third example (crosses) has white observations with a density closely resembling the Gaussian: a hyperbolic secant density having zero mean and unit variance. The sum of squared observations for each example are shown with the optimal χ^2_{100} interval displayed. Note how dramatically the test statistic departs from the decision interval when parameters disagree.

5.1.6 Stein's Lemma

As important as understanding the false-alarm, miss and error probabilities of the likelihood ratio test might be, no general expression exists for them. In analyzing the Gaussian problem, we find these in terms of $Q(\cdot)$, which has no closed form expression. In the general problem, the situation is much worse: No expression of any kind can be found! The reason is that we don't have the probability distribution of the sufficient statistic in the likelihood ratio test, which is needed in the expressions (5.8) {156}. We are faced with the curious situation that while knowing the decision rule that optimizes the performance probabilities, we usually don't know what the resulting performance probabilities will be.

Some general expressions are known for the *asymptotic* form of these error probabilities: limiting expressions as the number of observations L becomes large. These results go by the generic name of *Stein's Lemma* [23: §11.8].*

In the case that the L observations are statistically independent according to both models, Stein-Chernoff Lemma states that the false-alarm probability resulting from using the Neyman-Pearson hypothesis test has the form [54]

$$\lim_{L\to\infty} \frac{1}{L} \log P_F = -\mathscr{D}(p_{\mathbf{X}|\mathscr{M}_1} \| p_{\mathbf{X}|\mathscr{M}_0}) \quad \text{or} \quad P_F \stackrel{L\to\infty}{\longrightarrow} f(L) e^{-L\mathscr{D}(p_{\mathbf{X}|\mathscr{M}_1} \| p_{\mathbf{X}|\mathscr{M}_0})} , \tag{5.9}$$

where $\mathscr{D}(\cdot\|\cdot)$ is the Kullback-Leibler distance between two densities introduced in §2.4 {22}. $f(L)$ denotes a slowly varying function compared to the exponential: $\lim_{L\to\infty}[\log f(L)]/L = 0$. This function is problem- and P_M-dependent, and thus usually not known. What this asymptotic formula says is that as the number of observations increases, the false-alarm probability of a Neyman-Pearson hypothesis test plotted on semilogarithmic coordinates will eventually become a straight line, the slope of which is $-\mathscr{D}(p_1\|p_0)$. Figure 5.5 demonstrates this effect. A similar result holds for the miss probability if it is optimized and the false-alarm probability is constrained. Thus, in *all* model evaluation problems solved using the Neyman-Pearson approach, the opti-

*The attribution to statistician Charles Stein is probably incorrect. Herman Chernoff wrote a paper [22] providing a derivation of this result. A reviewer stated that he thought Stein had derived the result in a technical report, which Chernoff had not seen. Chernoff modified his paper to include the reference without checking it. Chernoff's paper provided the link to Stein. However, Stein later denied he had proved the result; so much for not questioning reviewers! Stein's Lemma should be known as Chernoff's Lemma.

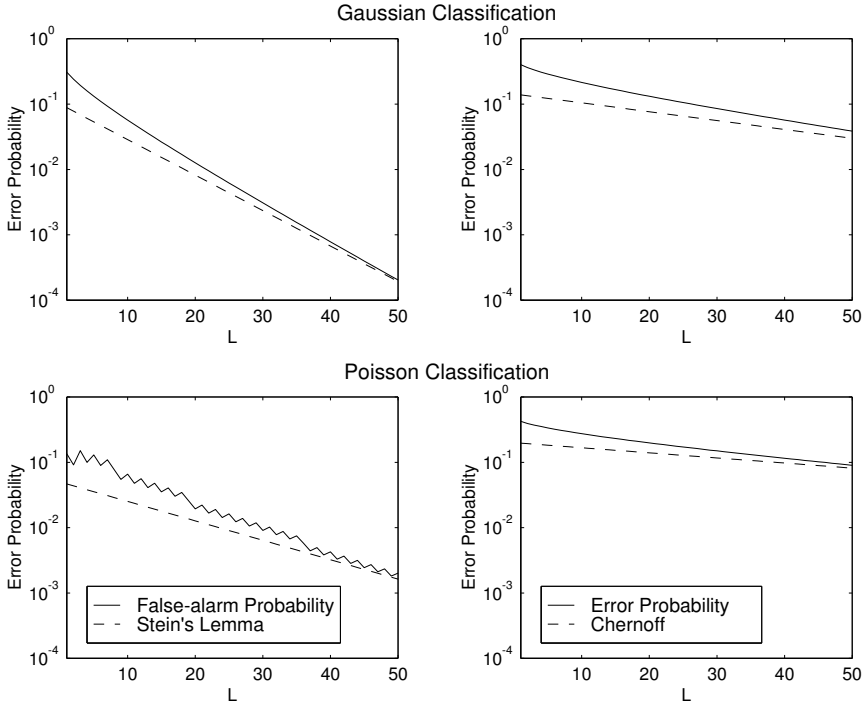


Figure 5.5: Using the Gaussian and Poisson classification problems as examples, we plot the false-alarm probability (left panels) and average probability of error (right panels) for each as a function of the amount of statistically independent data used by the optimal classifier. The miss-probability criterion was that it be less than or equal to 0.1. The *a priori* probabilities are $1/2$ in the right-column examples. As shown here, the average error probability produced by the minimum P_e classifier typically decays more slowly than the false-alarm probability for the classifier that fixes the miss probability. The dashed lines depict the behavior of the error probabilities as predicted by asymptotic theory (5.9). In each case, these theoretical lines have been shifted vertically for ease of comparison.

---

mized probability will always (eventually) decay exponentially in the number of observations. For this reason, $\mathscr{D}(p_{\mathbf{X}|\mathscr{M}_1} \| p_{\mathbf{X}|\mathscr{M}_0})$ is known as the *exponential rate*.

Showing this result relies on, believe it or not, the law of large numbers. Define the decision region $\mathfrak{R}_1(L)$ according to

$$\mathfrak{R}_1(L) = \left\{ \mathbf{X} \colon e^{L(\mathscr{D}(p_{X|\mathscr{M}_1} \| p_{X|\mathscr{M}_0}) - \delta)} \leq \frac{p_{\mathbf{X}|\mathscr{M}_1}}{p_{\mathbf{X}|\mathscr{M}_0}} \leq e^{L(\mathscr{D}(p_{X|\mathscr{M}_1} \| p_{X|\mathscr{M}_0}) + \delta)} \right\}$$

This decision region will vary with the number of observations as is typical of Neyman-Pearson decision rules. First of all, we must show that the decision region corresponding to this rule is equivalent to specifying a criterion on the miss probability.

$$1 - P_M = \Pr\left[\frac{1}{L} \sum_l \log \frac{p(X_l|\mathscr{M}_1)}{p(X_l|\mathscr{M}_0)} \in \left(\mathscr{D}(p_{\mathbf{X}|\mathscr{M}_1} \| p_{\mathbf{X}|\mathscr{M}_0}) - \delta, \mathscr{D}(p_{\mathbf{X}|\mathscr{M}_1} \| p_{\mathbf{X}|\mathscr{M}_0}) + \delta \right) \middle| \mathscr{M}_1 \right]$$

This sum is the average of the log-likelihood ratios computed at each observation assuming that $\mathscr{M}_1$ is true. Because of the strong law of large numbers, as the number of observations increases, this sum converges to its expected value, which is $\mathscr{D}(p_{\mathbf{X}|\mathscr{M}_1} \| p_{\mathbf{X}|\mathscr{M}_0})$. Therefore, $\lim_{L\to\infty} 1 - P_M = 1$ for any δ, which means $P_M \to 0$. Thus, this decision region guarantees not only that the miss probability is less than some specified number, but also that it decreases systematically as the number of observations increases.

To analyze how the false-alarm probability varies with L, we note that in this decision region

$$p_{\mathbf{X}|\mathcal{M}_0} \leq p_{\mathbf{X}|\mathcal{M}_1} e^{-L(\mathscr{D}(p_{X|\mathcal{M}_1}\|p_{X|\mathcal{M}_0})-\delta)}$$
$$p_{\mathbf{X}|\mathcal{M}_0} \geq p_{\mathbf{X}|\mathcal{M}_1} e^{-L(\mathscr{D}(p_{X|\mathcal{M}_1}\|p_{X|\mathcal{M}_0})+\delta)}$$

Integrating these over $\Re_1(L)$ yields upper and lower bounds on P_F.

$$e^{-L(\mathscr{D}(p_{X|\mathcal{M}_1}\|p_{X|\mathcal{M}_0})+\delta)}(1-P_M) \leq P_F \leq e^{-L(\mathscr{D}(p_{X|\mathcal{M}_1}\|p_{X|\mathcal{M}_0})-\delta)}(1-P_M)$$

or

$$-\mathscr{D}(p_{X|\mathcal{M}_1}\|p_{X|\mathcal{M}_0}) - \delta + \frac{1}{L}\log(1-P_M) \leq \frac{1}{L}\log P_F \leq -\mathscr{D}(p_{X|\mathcal{M}_1}\|p_{X|\mathcal{M}_0}) + \delta + \frac{1}{L}\log(1-P_M)$$

Because δ can be made arbitrarily small as the number of observations increases, we obtain Stein's Lemma. This derivation assumed a particular structure for the decision region; how do we know it is the optimal decision region? Its optimality arises because an auxiliary proof shows that no smaller exponent exists.

For the average probability of error, the Chernoff distance {22} is its exponential rate.

$$P_e \stackrel{L\to\infty}{\longrightarrow} f(L)e^{-L\mathscr{C}(p_{\mathbf{X}|\mathcal{M}_0}, p_{\mathbf{X}|\mathcal{M}_1})}$$

Showing this result leads to more interesting results. For the minimum-probability-of-error detector, the decision region $\Re_1$ is defined according to $\pi_1 p_{\mathbf{X}|\mathcal{M}_1} > \pi_0 p_{\mathbf{X}|\mathcal{M}_0}$. Using this fact, we can write the expression for P_e as

$$\begin{aligned} P_e &= \pi_0 \int_{\Re_1} p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{X}|\mathcal{M}_0)\, d\mathbf{X} + \pi_1 \int_{\Re_0} p_{\mathbf{X}|\mathcal{M}_1}(\mathbf{X}|\mathcal{M}_1)\, d\mathbf{X} \\ &= \int \min\{\pi_0 p_{\mathbf{X}|\mathcal{M}_0}, \pi_1 p_{\mathbf{X}|\mathcal{M}_1}\}\, d\mathbf{X} \end{aligned}$$

The minimum function has the property $\min\{a,b\} \leq a^{1-s}b^s$, $0 \leq s \leq 1$ for non-negative quantities a, b.* With this bound, we can find an upper bound for the average error probability.

$$\begin{aligned} P_e &\leq \int \left[\pi_0 p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{X}|\mathcal{M}_0)\right]^{1-s} \left[\pi_1 p_{\mathbf{X}|\mathcal{M}_1}(\mathbf{X}|\mathcal{M}_1)\right]^s d\mathbf{X} \\ &\leq \int \left[p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{X}|\mathcal{M}_0)\right]^{1-s} \left[p_{\mathbf{X}|\mathcal{M}_1}(\mathbf{X}|\mathcal{M}_1)\right]^s d\mathbf{X} \end{aligned}$$

When $\mathbf{X}$ has statistically independent and identically distributed components,

$$\begin{aligned} P_e &\leq \int \left[\prod_l p_{X_l|\mathcal{M}_0}(X_l|\mathcal{M}_0)\right]^{1-s} \left[\prod_l p_{X_l|\mathcal{M}_1}(X_l|\mathcal{M}_1)\right]^s d\mathbf{X} \\ &= \prod_l \int \left[p_{X_l|\mathcal{M}_0}(X_l|\mathcal{M}_0)\right]^{1-s} \left[p_{X_l|\mathcal{M}_1}(X_l|\mathcal{M}_1)\right]^s dX_l \\ &= \left\{\int \left[p_{X|\mathcal{M}_0}(X|\mathcal{M}_0)\right]^{1-s} \left[p_{X|\mathcal{M}_1}(X|\mathcal{M}_1)\right]^s dX\right\}^L \end{aligned}$$

Consequently, we have that

$$\frac{1}{L}\log P_e \leq \log \int \left[p_{X|\mathcal{M}_0}(X|\mathcal{M}_0)\right]^{1-s} \left[p_{X|\mathcal{M}_1}(X|\mathcal{M}_1)\right]^s dX, \quad \text{for all } s\,.$$

Asymptotically, this bound is tight (Gaussian example shows this to be true). Thus, the exponential rate for the average probability of error is the minimum of this bound with respect to s. This quantity is the negative

*To see this, simply plot the minimum function versus one of its arguments with the other one fixed. Plotting $a^{1-s}b^s$ the same way shows that it indeed is an upper bound.

of the Chernoff distance (2.10) and we arrive at the expression for the average probability of error. Fig. 5.5 shows that the exponential rates of minimum P_e and Neyman-Pearson hypothesis tests are not necessarily the same. In other words, the distance to the equi-distant distribution need not equal the total distance, which seems to make sense.

The larger the distance, the greater the rate at which the error probability decreases, which corresponds to an easier problem (smaller error probabilities). However, Stein-Chernoff Lemma does not allow a precise calculation of error probabilities. The quantity $f(\cdot)$ depends heavily on the underlying probability distributions in ways that are problem dependent. All a distance calculation gives us is the exponential rate.

Example

For the Gaussian example we have been considering, set the miss probability's criterion value to be α. Because $P_M = Q\left(\frac{Lm-\gamma}{\sqrt{L}\sigma}\right)$, we find the threshold value γ to be $Lm - \sqrt{L}\sigma Q^{-1}(\alpha)$, which yields a false-alarm probability of $Q\left(\frac{\sqrt{L}m}{\sigma} - Q^{-1}(\alpha)\right)$. The Kullback-Leibler distance between two Gaussians $p_0 = \mathcal{N}(0,\sigma^2)$ and $p_1 = \mathcal{N}(m,\sigma^2)$ equals

$$\begin{aligned}\mathscr{D}(p_1\|p_0) &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} \cdot \left[-\frac{(x-m)^2}{2\sigma^2} + \frac{x^2}{2\sigma^2}\right] dx \\ &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} \cdot \left[\frac{2mx-m^2}{2\sigma^2}\right] dx \\ &= \frac{2m\cdot m - m^2}{2\sigma^2} \\ &= \frac{m^2}{2\sigma^2}\end{aligned}$$

In this Gaussian case the Kullback-Leibler distance is symmetric: $\mathscr{D}(p_1\|p_0) = \mathscr{D}(p_0\|p_1)$. Verify this atypical (most cases are asymmetric) result for yourself. The Chernoff distance equals $m^2/8\sigma^2$ (the value of s^* is $\frac{1}{2}$). Stein's Lemma predicts that the false-alarm probability has the asymptotic form $f(L)\exp\{-Lm^2/2\sigma^2\}$. Thus, does $Q\left(\frac{\sqrt{L}m}{\sigma} - Q^{-1}(\alpha)\right) \to f(L)\exp\{-Lm^2/2\sigma^2\}$, where $f(L)$ is slowly varying compare to the exponential? An asymptotic formula (equation (2.2) {8}) for $Q(x)$ is

$$Q(x) \xrightarrow{x\to\infty} \frac{1}{\sqrt{2\pi}x} e^{-x^2/2} .$$

As L increases, so does the argument of $Q(\cdot)$ in the expression for the false-alarm probability. Thus

$$\begin{aligned}P_F &\xrightarrow{L\to\infty} \frac{1}{\sqrt{2\pi}\left(\sqrt{L}m/\sigma - Q^{-1}(\alpha)\right)} \exp\left\{-\left(\frac{\sqrt{L}m}{\sigma} - Q^{-1}(\alpha)\right)^2/2\right\} \\ &= \frac{1}{\sqrt{2\pi}\left(\sqrt{L}m/\sigma - Q^{-1}(\alpha)\right)} \exp\left\{\frac{\sqrt{L}m}{\sigma}Q^{-1}(\alpha) - [Q^{-1}(\alpha)]^2/2\right\} \cdot e^{-Lm^2/2\sigma^2}\end{aligned}$$

The quantity multiplying the final exponential corresponds to $f(L)$, which satisfies the slowly varying property. We have verified Stein's Lemma prediction for the asymptotic behavior of the false-alarm probability in the Gaussian case. Note how the criterion value α occurs only in the expression for $f(L)$, and hence does *not* affect the exponential rate; Stein's Lemma says this situation always occurs.

5.2 Sequential Hypothesis Testing

In many circumstances, the observations to be used in evaluating models arrive sequentially rather than all at once. For example, you may not want to wait until all the data become available to make a time-critical decision. The decision rules we have derived implicitly assume the entire *block* of data is available. You might wonder whether a hypothesis test could be developed that takes the sequential arrival of data into account, making decisions as the data arrive, with the possibility of determining *early* in the data collection procedure the validity of one model, while maintaining the *same* performance specifications. Answering this question leads to the formulation of *sequential hypothesis testing* [80: pp. 136–156],[90]. Not only do *sequential* tests exist, they can provide performance superior to that of block tests in certain cases.

To make decisions as the data become available, we must generalize the decision-making process. Assume as before that the observed data comprise an observation vector $\mathbf{X}$ of length L. The decision rule (in the two-model case) now consists of determining which model is valid *or* that more data are required. Thus, the range of values of $\mathbf{X}$ is partitioned into three regions $\Re_0$, $\Re_1$, and $\Re_?$. Making the latter decision implies that the data gathered to that point are insufficient to meet the performance requirements. More data must be obtained to achieve the required performance and the test re-applied once these additional data become available. Thus, a *variable* number of observations are required to make a decision. An issue in this kind of procedure is the number of observations required to satisfy the performance criteria: for a common set of performance specifications, does this procedure result in a decision rule requiring, on the average, fewer observations than does a fixed-length block test?

5.2.1 Sequential Likelihood Ratio Test

In a manner similar to the Neyman-Pearson criterion, we specify the false-alarm probability P_F; in addition, we need to specify the detection probability P_D. These constraints over-specify the model evaluation problem where the number of observations is fixed: enforcing one constraint forces violation of the other. In contrast, both may be specified the sequential test as we shall see.

Assuming a likelihood ratio test, two thresholds are required to define to the three decision regions.

$$\begin{array}{rcll} & \Lambda_L(\mathbf{X}) & < \eta_0 & \text{say } \mathscr{M}_0 \\ \eta_0 < & \Lambda_L(\mathbf{X}) & < \eta_1 & \text{say ``need more data''} \\ \eta_1 < & \Lambda_L(\mathbf{X}) & & \text{say } \mathscr{M}_1\,, \end{array}$$

where $\Lambda_L(\mathbf{X})$ is the usual likelihood ratio where the dimension L of the vector $\mathbf{X}$ is explicitly denoted. The threshold values η_0 and η_1 are found from the constraints, which are expressed as

$$P_F = \int_{\Re_1} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)\,d\mathbf{X} = \alpha \quad \text{and} \quad P_D = \int_{\Re_1} p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)\,d\mathbf{X} = \beta\,.$$

Here, α and β are design constants that you choose according to the application. Note that the probabilities P_F, P_D are associated *not* with what happens on a given trial, but what the sequential test yields in terms of performance when a decision is made. Thus, $P_M = 1 - P_D$, but the probability of correctly saying $\mathscr{M}_1$ on a given trial does not equal one minus the probability of incorrectly saying $\mathscr{M}_0$ is true: The "need more data" region must be accounted for on an individual trial but not when considering the sequential test's performance when it terminates.

Rather explicitly attempting to relate thresholds to performance probabilities, we obtain simpler results by using bounds and approximations. Note that the expression for P_D may be written as

$$\begin{aligned} P_D &= \int_{\Re_1} \frac{p_{\mathbf{x}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1)}{p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)}\, p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)\,d\mathbf{X} \\ &= \int_{\Re_1} \Lambda_L(\mathbf{X})\, p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)\,d\mathbf{X}\,. \end{aligned}$$

In the decision region $\Re_1$, $\Lambda_L(\mathbf{X}) \geq \eta_1$; thus, a lower bound on the detection probability can be established by substituting this inequality into the integral.

$$P_D \geq \eta_1 \int_{\Re_1} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)\,d\mathbf{X}$$

The integral is the false-alarm probability P_F of the test when it terminates. In this way, we find that $P_D/P_F \geq \eta_1$. Using similar arguments on the miss probability, we obtain a similar bound on the threshold η_0. These inequalities are summarized as

$$\eta_0 \geq \frac{1-P_D}{1-P_F} \quad \text{and} \quad \eta_1 \leq \frac{P_D}{P_F}.$$

These bounds, which relate the thresholds in the sequential likelihood ratio test with the false-alarm and detection probabilities, are general, applying even when sequential tests are not being used. In the usual likelihood ratio test, there is a single threshold η; these bounds apply to it as well, implying that in a likelihood ratio test the error probabilities will *always* satisfy

$$\boxed{\frac{P_D}{P_F} \geq \frac{1-P_D}{1-P_F}}. \tag{5.10}$$

This relationship can be manipulated to show $P_D \geq P_F$, indicating that the likelihood ratio test is, at the very least, a reasonable decision rule and that the ROC curves of §5.1.3 {156} have the right general form.

Only with difficulty can we solve the inequality constraints on the sequential test's thresholds in the general case. Surprisingly, by approximating the inequality constraints by *equality* constraints we can obtain a result having pragmatically important properties. As an approximation, we thus turn to solving for η_0 and η_1 under the conditions

$$\eta_0 = \frac{1-\beta}{1-\alpha} \quad \text{and} \quad \eta_1 = \frac{\beta}{\alpha}.$$

In this way, the threshold values are explicitly specified in terms of the desired performance probabilities. We use the criterion values for the false-alarm and detection probabilities because when we use these equalities, the test's resulting performance probabilities P_F and P_D usually do *not* satisfy the design criteria. For example, equating η_1 to a value potentially larger than its desired value might result in a smaller detection probability and a larger false-alarm rate. We will want to understand how much actual performance departs from what we want.

The relationships derived above between the performance levels and the thresholds apply no matter how the thresholds are chosen.

$$\frac{1-\beta}{1-\alpha} \geq \frac{1-P_D}{1-P_F} \qquad \frac{\beta}{\alpha} \leq \frac{P_D}{P_F} \tag{5.11}$$

From these inequalities, two important results follow:

$$\frac{1-\beta}{1-\alpha} \geq 1-P_D \quad \frac{\alpha}{\beta} \geq P_F \qquad \text{and} \qquad P_F + (1-P_D) \leq \alpha + (1-\beta).$$

The first result follows directly from the threshold bounds. To derive the second result, we must work a little harder. Multiplying the first inequality in (5.11) by $(1-\alpha)(1-P_F)$ yields $(1-\beta)(1-P_F) \geq (1-\alpha)(1-P_D)$. Considering the reciprocal of the second inequality in (5.11) and multiplying it by βP_D yields $\alpha P_D \geq \beta P_F$. Adding these two inequalities yields the second result.

The first set of inequalities suggest that the false-alarm and miss (which equals $1-P_D$) probabilities will increase only slightly from their specified values: the denominators on the right sides are very close to unity in the interesting cases (e.g., small error probabilities like 0.01). The second inequality suggests that the sum of the false-alarm and miss probabilities obtained in practice will be less than the sum of the specified error probabilities. Taking these results together, one of two situations will occur when we approximate the inequality criterion by equality: either the false alarm probability will decrease and the detection probability increase (a most pleasing but unlikely circumstance) *or* one of the error probabilities will increase while the other decreases. *The false-alarm and miss probabilities cannot both increase.* Furthermore, whichever one increases, the first inequalities suggest that the incremental change will be small. Our *ad hoc* approximation to the thresholds does indeed yield a level of performance close to that specified.

Usually, the likelihood is manipulated to derive a sufficient statistic, resulting in the sequential decision rule

$$\begin{array}{lll} \Upsilon_L(\mathbf{X}) & < \gamma_0(L) & \text{say } \mathscr{M}_0 \\ \gamma_0(L) < \Upsilon_L(\mathbf{X}) & < \gamma_1(L) & \text{say "need more data"} \\ \gamma_1(L) < \Upsilon_L(\mathbf{X}) & & \text{say } \mathscr{M}_1 . \end{array}$$

Note that the thresholds $\gamma_0(L)$ and $\gamma_1(L)$, derived from the thresholds η_0 and η_1, usually depend on the number of observations used in the decision rule.

Example

Let $\mathbf{X}$ be a Gaussian random vector as in our previous examples with statistically independent components.

$$\mathscr{M}_0 : \mathbf{X} \sim \mathscr{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$
$$\mathscr{M}_1 : \mathbf{X} \sim \mathscr{N}(\mathbf{m}, \sigma^2 \mathbf{I})$$

The mean vector $\mathbf{m}$ is assumed for simplicity to consist of equal positive values: $\mathbf{m} = \text{col}[m, \ldots, m]$, $m > 0$. Using the previous derivations, our sequential test becomes

$$\begin{array}{lll} \displaystyle\sum_{l=0}^{L-1} X_l & < \dfrac{\sigma^2}{m} \ln \eta_0 + \dfrac{Lm}{2} & \text{say } \mathscr{M}_0 \\ \dfrac{\sigma^2}{m} \ln \eta_0 + \dfrac{Lm}{2} < \displaystyle\sum_{l=0}^{L-1} X_l & < \dfrac{\sigma^2}{m} \ln \eta_1 + \dfrac{Lm}{2} & \text{say "need more data"} \\ \dfrac{\sigma^2}{m} \ln \eta_1 + \dfrac{Lm}{2} < \displaystyle\sum_{l=0}^{L-1} X_l & & \text{say } \mathscr{M}_1 . \end{array}$$

Starting with $L = 1$, we gather the data and compute the sum. The sufficient statistic will lie in the middle range between the two thresholds until one of them is exceeded as shown in Fig. 5.6. The model evaluation procedure then terminates and the chosen model announced. Note how the thresholds depend on the amount of data available (as expressed by L). This variation typifies the sequential hypothesis tests.

5.2.2 Average Number of Required Observations

The awake reader might wonder whether that the sequential likelihood ratio test just derived has the disturbing property that it may never terminate: can the likelihood ratio wander between the two thresholds forever? Fortunately, the sequential likelihood ratio test has been shown to terminate with probability one [90]. Confident of eventual termination, we need to explore how many observations are required to meet performance specifications. The number of observations is variable, depending on the observed data and the stringency of the specifications. The *average* number of observations required can be determined in the interesting case when the observations are statistically independent.

Assuming that the observations are statistically independent and identically distributed, the likelihood ratio is equal to the product of the likelihood ratios evaluated at each observation. Considering $\ln \Lambda_{L_0}(\mathbf{X})$, the logarithm of the likelihood ratio when a decision is made on observation L_0, we have

$$\ln \Lambda_{L_0}(\mathbf{X}) = \sum_{l=0}^{L_0-1} \ln \Lambda(X_l) ,$$

where $\Lambda(X_l)$ is the likelihood ratio corresponding to the l^{th} observation. We seek an expression for $\mathsf{E}[L_0]$, the expected value of the number of observations required to make the decision. To derive this quantity, we evaluate the expected value of the likelihood ratio when the decision is made. This value will usually vary

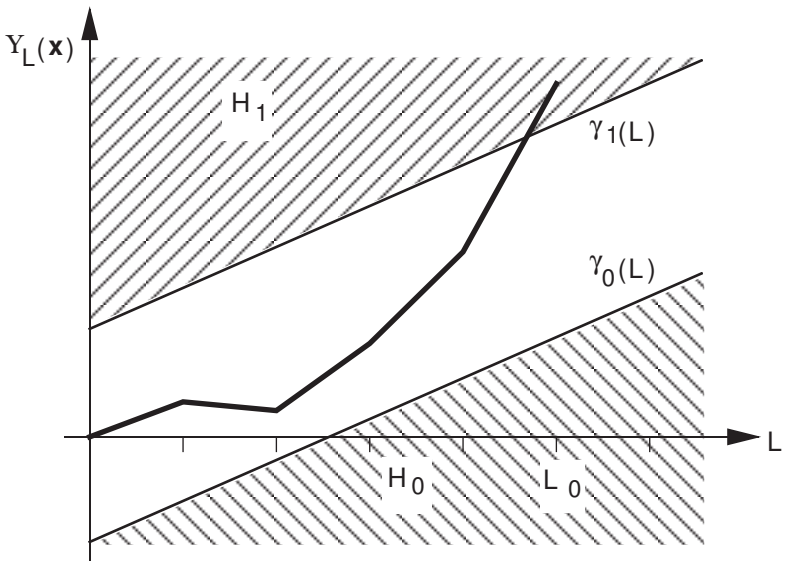


Figure 5.6: The sufficient statistic wanders between the two thresholds in a sequential decision rule until one of them is crossed by the statistic. The number of observations used to obtain a decision is L_0.

with which model is actually valid; we must consider both models separately. Using the laws of conditional expectation (see §2.1.5 {7}), we find that the expected value of $\ln \Lambda_{L_0}(\mathbf{X})$, assuming that model $\mathscr{M}_1$ was true, is given by

$$\mathsf{E}[\ln \Lambda_{L_0}(\mathbf{X})|\mathscr{M}_1] = \mathsf{E}\big[\mathsf{E}[\ln \Lambda_{L_0}(\mathbf{X})|\mathscr{M}_1, L_0]\big]$$

The outer expected value is evaluated with respect to the probability distribution of L_0; the inner expected value is average value of the log-likelihood assuming that L_0 observations were required to choose model $\mathscr{M}_1$. In the latter case, the log-likelihood is the sum of L_0 component log-likelihood ratios

$$\mathsf{E}[\ln \Lambda_{L_0}(\mathbf{X})|\mathscr{M}_1, L_0] = L_0\, \mathsf{E}[\ln \Lambda_1(X)|\mathscr{M}_1]$$

Noting that the expected value on the right is a constant with respect to the outer expected value, we find that

$$\mathsf{E}[\ln \Lambda_{L_0}(\mathbf{X})|\mathscr{M}_1] = \mathsf{E}[L_0|\mathscr{M}_1] \cdot \mathsf{E}[\ln \Lambda_1(X)|\mathscr{M}_1]\,.$$

The average number of observations required to make a decision, correct or incorrect, assuming that $\mathscr{M}_1$ is true is thus expressed by

$$\boxed{\mathsf{E}[L_0|\mathscr{M}_1] = \frac{\mathsf{E}[\ln \Lambda_{L_0}(\mathbf{X})|\mathscr{M}_1]}{\mathsf{E}[\ln \Lambda_1(X)|\mathscr{M}_1]}}\,.$$

Assuming that the other model was true, we have the complementary result

$$\boxed{\mathsf{E}[L_0|\mathscr{M}_0] = \frac{\mathsf{E}[\ln \Lambda_{L_0}(\mathbf{X})|\mathscr{M}_0]}{\mathsf{E}[\ln \Lambda_1(X)|\mathscr{M}_0]}}\,.$$

The numerator is difficult to calculate exactly but easily approximated; assuming that the likelihood ratio equals its threshold value when the decision is made,

$$\begin{aligned}
\mathsf{E}[\ln \Lambda_{L_0}(\mathbf{X})|\mathscr{M}_0] &\approx P_F \ln \eta_1 + (1-P_F)\ln \eta_0 \\
&= P_F \ln\left(\frac{P_D}{P_F}\right) + (1-P_F)\ln\left(\frac{1-P_D}{1-P_F}\right), \\
\mathsf{E}[\ln \Lambda_{L_0}(\mathbf{X})|\mathscr{M}_1] &\approx P_D \ln \eta_1 + (1-P_D)\ln \eta_0 \\
&= P_D \ln\left(\frac{P_D}{P_F}\right) + (1-P_D)\ln\left(\frac{1-P_D}{1-P_F}\right).
\end{aligned}$$

Note these expressions are *not* problem dependent; they depend only on the specified probabilities. The denominator cannot be approximated in a similar way with such generality; it must be evaluated for each problem.

Example

In the Gaussian example we have been exploring, the log-likelihood of each component observation X_l is given by

$$\ln \Lambda(X_l) = \frac{mX_l}{\sigma^2} - \frac{m^2}{2\sigma^2}$$

The conditional expected values required to evaluate the expression for the average number of required observations are

$$\mathsf{E}[\ln \Lambda_1(X)|\mathscr{M}_0] = -\frac{m^2}{2\sigma^2} \qquad \mathsf{E}[\ln \Lambda_1(X)|\mathscr{M}_1] = \frac{m^2}{2\sigma^2}.$$

For simplicity, let's assume that the false-alarm and detection probabilities are symmetric (i.e., $P_F = 1 - P_D$). The expressions for the average number of observations are equal for each model and we have

$$\mathsf{E}[L_0|\mathscr{M}_0] = \mathsf{E}[L_0|\mathscr{M}_1] = f(P_F)\frac{\sigma^2}{m^2},$$

where $f(P_F)$ is a function equal to $(2-4P_F)\ln\left(\frac{1-P_F}{P_F}\right)$. Thus, the number of observations decreases with increasing signal-to-noise ratio m/σ and increases as the false-alarm probability is reduced.

Suppose we used a likelihood ratio test where all data were considered once and a decision made; how many observations would be required to achieve a specified level of performance and how would this fixed number compare with the average number of observations in a sequential test? In this example, we find from our earlier calculations {156} that $P_F = Q\left(\frac{\sqrt{L}m}{2\sigma}\right)$ so that

$$L = 4\left[Q^{-1}(P_F)\right]^2 \frac{\sigma^2}{m^2}.$$

The duration of the sequential and block tests depend on the signal-to-noise ratio in the same way; however, the dependence on the false-alarm probability is quite different. As depicted in Fig. 5.7, the disparity between these quantities increases rapidly as the false alarm probability decreases, with the sequential test requiring correspondingly fewer observations *on the average*.

We must not forget that these results apply to the *average* number of observations required to make a decision. Expressions for the distribution of the number of observations are complicated and depend heavily on the problem. When an extremely large number of observation are required to resolve a difficult case to the required accuracy, we are forced to *truncate* the sequential test, stopping when a specified number of observations have been used. A decision would then be made by dividing the region between the boundaries in half and selecting the model corresponding to the boundary nearest to the sufficient statistic. If this truncation point is larger than the expected number, the performance probabilities will change little. "Larger" is again problem dependent; analytic results are few, leaving the option of computer simulations to estimate the distribution of the number of observations required for a decision.

5.3 Detection in the Presence of Unknowns

We assumed in the previous sections that we have a few well-specified models (hypotheses) for a set of observations. These models were probabilistic; to apply the techniques of statistical hypothesis testing, the models take the form of conditional probability densities. In many interesting circumstances, the exact nature of these densities may not be known. For example, we may know *a priori* that the mean is either zero or some constant (as in the Gaussian example). However, the variance of the observations may not be known or the value of the non-zero mean may be in doubt. In a signal processing context, these respective situations could occur when the background noise level is unknown (a likely possibility in applications) or when the

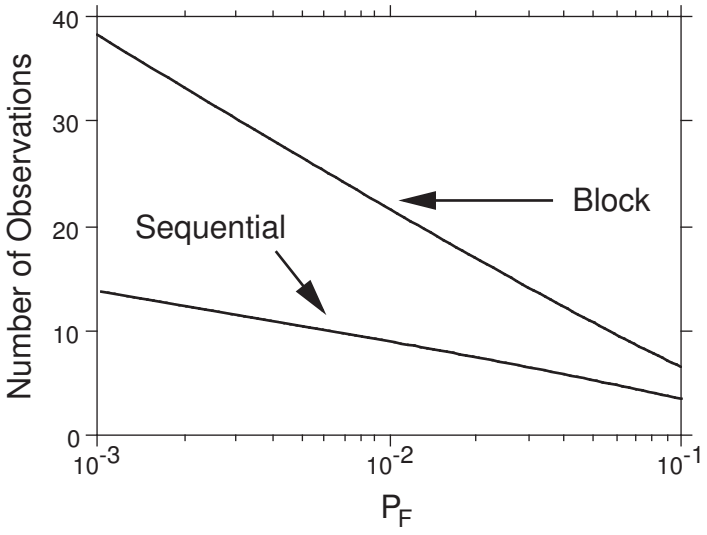


Figure 5.7: The numbers of observations required by the sequential test (on the average) and by the block test for Gaussian observations are proportional to σ^2/m^2; the coefficients of these expressions ($f(P_F)$ and $4\left[Q^{-1}(P_F)\right]^2$ respectively) are shown.

signal amplitude is not known because of far-field range uncertainties (the further the source of propagating energy, the smaller its received energy at each sensor). In an extreme case, we can question the exact nature of the probability densities (everything is not necessarily Gaussian!). The model evaluation problem can still be posed for these situations; we classify the "unknown" aspects of a model testing problem as either *parametric* (the variance is not known, for example) or *nonparametric* (the formula for the density is in doubt). The former situation has a relatively long history compared to the latter; many techniques can be used to approach parametric problems while the latter is a subject of current research [38]. We concentrate on parametric problems here.

We describe the dependence of the conditional density on a set of parameters by incorporating the parameter vector $\boldsymbol{\theta}$ as part of the condition. We write the likelihood function as $p_{\mathbf{X}|\mathscr{M}_i,\boldsymbol{\theta}}(\mathbf{x}|\mathscr{M}_i,\boldsymbol{\theta})$ for the parametric problem. In statistics, this situation is said to be a *composite hypothesis* [25: p. 528]. Such situations can be further categorized according to whether the parameters are *random* or *nonrandom*. For a parameter to be random, we have an expression for its *a priori* density, which could depend on the particular model. As stated many times, a specification of a density usually expresses some knowledge about the range of values a parameter may assume *and* the relative probability of those values. Saying that a parameter has a uniform distribution implies that the values it assumes *are* equally likely, *not* that we have no idea what the values might be and express this ignorance by a uniform distribution. If we are ignorant of the underlying probability distribution that describes the values of a parameter, we will characterize them simply as being *unknown* (not random). Once we have considered the random parameter case, nonrandom but unknown parameters will be discussed.

5.3.1 Random Parameters

When we know the density of $\boldsymbol{\theta}$, the likelihood function can be expressed as

$$p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{x}|\mathscr{M}_i) = \int p_{\mathbf{X}|\mathscr{M}_i,\boldsymbol{\theta}}(\mathbf{x}|\mathscr{M}_i,\boldsymbol{\theta})\, p_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathscr{M}_i)\, d\boldsymbol{\theta}$$

and the likelihood ratio in the random parameter case becomes

$$\boxed{\Lambda(\mathbf{X}) = \frac{\int p_{\mathbf{X}|\mathscr{M}_1,\boldsymbol{\theta}}(\mathbf{x}|\mathscr{M}_1,\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathscr{M}_1)\, d\boldsymbol{\theta}}{\int p_{\mathbf{X}|\mathscr{M}_0,\boldsymbol{\theta}}(\mathbf{x}|\mathscr{M}_0,\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathscr{M}_0)\, d\boldsymbol{\theta}}}.$$

Unfortunately, there are many examples where either the integrals involved are intractable or the sufficient statistic is virtually the same as the likelihood ratio, which can be difficult to compute.

Example

A simple, but interesting, example that results in a computable answer occurs when the mean of Gaussian random variables is either zero (model 0) or is $\pm m$ with equal probability (hypothesis 1). The second hypothesis means that a non-zero mean is present, but its sign is not known. We are therefore stating that if hypothesis one is in fact valid, the mean has fixed sign for each observation—what is random is its *a priori* value. As before, L statistically independent observations are made.

$$\begin{aligned} \mathscr{M}_0 &: \quad \mathbf{X} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathscr{M}_1 &: \quad \mathbf{X} \sim \mathscr{N}(\mathbf{m}, \sigma^2\mathbf{I}) \quad \mathbf{m} = \begin{cases} m\mathbf{1} & \text{Prob} = 1/2 \\ -m\mathbf{1} & \text{Prob} = 1/2 \end{cases} \end{aligned}$$

The numerator of the likelihood ratio is the sum of two Gaussian densities weighted by $1/2$ (the *a priori* probability values), one having a positive mean, the other negative. The likelihood ratio, after simple cancellation of common terms, becomes

$$\Lambda(\mathbf{X}) = \frac{1}{2}\exp\left\{\frac{2m\left(\sum_{l=0}^{L-1} X_l\right) - Lm^2}{2\sigma^2}\right\} + \frac{1}{2}\exp\left\{\frac{-2m\left(\sum_{l=0}^{L-1} X_l\right) - Lm^2}{2\sigma^2}\right\}$$

and the decision rule takes the form

$$\cosh\left(\frac{m}{\sigma^2}\sum_{l=0}^{L-1} X_l\right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \eta \exp\left\{\frac{Lm^2}{2\sigma^2}\right\},$$

where $\cosh(x)$ is the *hyperbolic cosine* given simply as $(e^x + e^{-x})/2$. As this quantity is an even function, the sign of its argument has no effect on the result. The decision rule can be written more simply as

$$\left|\sum_{l=0}^{L-1} X_l\right| \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\sigma^2}{|m|}\cosh^{-1}\left[\eta \exp\left\{\frac{Lm^2}{2\sigma^2}\right\}\right].$$

The sufficient statistic equals the *magnitude* of the sum of the observations in this case. While the right side of this expression, which equals γ, is complicated, it need only be computed once. Calculation of the performance probabilities can be complicated; in this case, the false-alarm probability is easy to find and the others more difficult.

5.3.2 Non-Random Parameters

In those cases where a probability density for the parameters cannot be assigned, the model evaluation problem can be solved in several ways; the methods used depend on the form of the likelihood ratio and the way in which the parameter(s) enter the problem. In the Gaussian problem we have discussed so often, the threshold used in the likelihood ratio test η may be unity. In this case, examination of the resulting computations required reveals that implementing the test *does not require knowledge of the variance of the observations* (see Problem 5.37). Thus, if the common variance of the underlying Gaussian distributions is not known, this lack of knowledge has *no effect* on the optimum decision rule. This happy situation—knowledge of the value of a parameter is not required by the optimum decision rule—occurs rarely, but should be checked before using more complicated procedures.

A second fortuitous situation occurs when the sufficient statistic as well as its probability density under one of the models do *not* depend on the unknown parameter(s). Although the sufficient statistic's threshold γ expressed in terms of the likelihood ratio's threshold η depends on the unknown parameters, γ may be computed as a single value using the Neyman-Pearson criterion *if the computation of the false-alarm probability does not involve the unknown parameters*.

Example

Continuing the example of the previous section, let's consider the situation where the value of the mean of each observation under model $\mathscr{M}_1$ is not known. The sufficient statistic is the sum of the observations (that quantity doesn't depend on m) and the distribution of the observation vector under model $\mathscr{M}_0$ does not depend on m (allowing computation of the false-alarm probability). However, a subtlety emerges; in the derivation of the sufficient statistic, we had to divide by the value of the mean. The critical step occurs once the logarithm of the likelihood ratio is manipulated to obtain

$$m\sum_{l=0}^{L-1} X_l \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta + \frac{Lm^2}{2}.$$

Recall that only *positively* monotonic transformations can be applied; if a negatively monotonic operation is applied to this inequality (such as multiplying both sides by -1), the *inequality reverses*. If the sign of m is known, it can be taken into account explicitly and a sufficient statistic results. If, however, the sign is not known, the above expression cannot be manipulated further and the left side constitutes the sufficient statistic for this problem. The sufficient statistic then depends on the unknown parameter and we cannot develop a decision rule in this case. If the sign is known, we can proceed. Assuming the sign of m is positive, the sufficient statistic is the sum of the observations $-\sum_{l=0}^{L-1} X_l$ — with the threshold γ found by

$$\gamma = \sqrt{L}\sigma Q^{-1}(P_F).$$

Note that if the variance σ^2 instead of the mean were unknown, we could not compute the threshold. The difficulty lies not with the sufficient statistic (it doesn't depend on the variance), but with calculating the false alarm probability: the probability density needed to find $P_F - p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0)$ — depends on the unknown variance. Another approach is required to deal with the unknown-variance problem.

When this situation occurs—the sufficient statistic *and* the false-alarm probability can be computed without needing the parameter in question, we have established what is known as a *uniformly most powerful test* (or UMP test) [25: pp. 529–531],[89: p. 89ff]. If an UMP test does not exist, which can only be demonstrated by explicitly finding the sufficient statistic and evaluating its probability distribution, then the composite hypothesis testing problem cannot be solved without some value for the parameter being used.

This seemingly impossible situation—we need the value of parameter that is assumed unknown—can be approached by noting that some data is available for "guessing" the value of the parameter. If a reasonable guess could be obtained, it could then be used in our model evaluation procedures developed in this chapter. *The data available for estimating unknown parameters are precisely the data used in the decision rule*. Procedures intended to yield "good" guesses of the value of a parameter are said to be *parameter estimates*. Estimation procedures are the topic of the next chapter; there we will explore a variety of estimation techniques and develop measures of estimate quality. For the moment, these issues are secondary; even if we knew the size of the estimation error, for example, the more pertinent issue is how the imprecise parameter value affects the performance probabilities. We can compute these probabilities *without* explicitly determining the estimate's error characteristics.

One parameter estimation procedure that fits nicely into the composite hypothesis testing problem is the *maximum likelihood estimate*.* Letting $\mathbf{X}$ denote the vector of observables and $\boldsymbol{\theta}$ a vector of parameters, the maximum likelihood estimate of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$, is that value of $\boldsymbol{\theta}$ that maximizes the conditional density $p_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{X}|\boldsymbol{\theta})$

*The maximum likelihood estimation procedure and its characteristics are fully described in §4.2.3 {68}.

of the observations given the parameter values. To use $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ in our decision rule, we estimate the parameter vector *separately* for each model, use the estimated value in the conditional density of the observations, and compute the likelihood ratio. This procedure is termed the *generalized likelihood ratio test* for the unknown parameter problem in hypothesis testing [65: p. 16],[89: p. 92ff].

$$\boxed{\Lambda(\mathbf{X}) = \frac{\max\limits_{\boldsymbol{\theta}} p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1;\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta}} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0;\boldsymbol{\theta})}} \tag{5.12}$$

Note that we do *not* find that value of the parameter that (necessarily) maximizes the likelihood ratio. Rather, we estimate the parameter value most consistent with the observed data in the context of each assumed model (hypothesis) of data generation. In this way, the estimate conforms with each potential model rather than being determined by some amalgam of supposedly mutually exclusive models.

Example

Returning to our Gaussian example, assume that the variance σ^2 is known but that the mean under $\mathscr{M}_1$ is unknown.

$$\begin{aligned} \mathscr{M}_0: &\quad \mathbf{X} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathscr{M}_1: &\quad \mathbf{X} \sim \mathscr{N}(m\mathbf{1}, \sigma^2\mathbf{I}) \quad m = ? \end{aligned}$$

The unknown quantity occurs only in the exponent of the conditional density under $\mathscr{M}_1$; to maximize this density, we need only to maximize the exponent. Thus, we consider the derivative of the exponent with respect to m.

$$\begin{aligned} \frac{\partial}{\partial m}\left[-\frac{1}{2\sigma^2}\sum_{l=0}^{L-1}(X_l - m)^2\right]\Bigg|_{m=\widehat{m}_{\mathrm{ML}}} &= 0 \\ \Longrightarrow \quad \sum_{l=0}^{L-1}(X_l - \widehat{m}_{\mathrm{ML}}) &= 0 \end{aligned}$$

The solution of this equation is the average value of the observations.

$$\widehat{m}_{\mathrm{ML}} = \frac{1}{L}\sum_{l=0}^{L-1} X_l$$

To derive the decision rule, we substitute this estimate in the conditional density for $\mathscr{M}_1$. The critical term, the exponent of this density, is manipulated to obtain

$$-\frac{1}{2\sigma^2}\sum_{l=0}^{L-1}\left(X_l - \frac{1}{L}\sum_{k=0}^{L-1} X_k\right)^2 = -\frac{1}{2\sigma^2}\left[\sum_{l=0}^{L-1} X_l^2 - \frac{1}{L}\left(\sum_{l=0}^{L-1} X_l\right)^2\right].$$

Noting that the first term in this exponent is identical to the exponent of the denominator in the likelihood ratio, the generalized likelihood ratio becomes

$$\Lambda(\mathbf{X}) = \exp\left\{+\frac{1}{2L\sigma^2}\left(\sum_{l=0}^{L-1} X_l\right)^2\right\}.$$

The sufficient statistic thus becomes the square (or equivalently the magnitude) of the summed observations. Compare this result with that obtained in the example {172}. There, an UMP test existed *if* we knew the sign of m and the sufficient statistic was the sum of the observations. Here, where we employed the generalized likelihood ratio test, we made no such assumptions about m; this generality accounts for the difference in sufficient statistic. Which test do you think would lead to a greater detection probability for a given false-alarm probability?

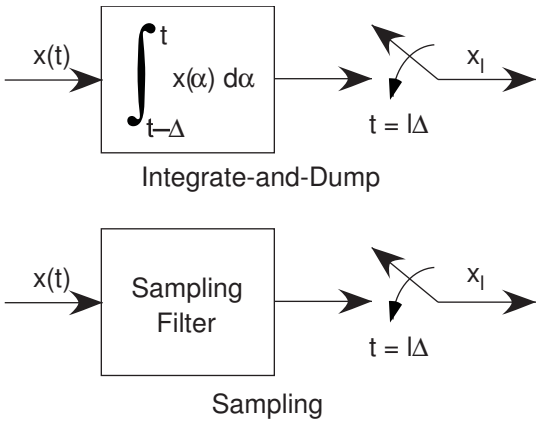


Figure 5.8: The two most common methods of converting continuous-time observations into discrete-time ones are shown. In the upper panel, the integrate-and-dump method is shown: the input is integrated over an interval of duration Δ and the result sampled. In the lower panel, the sampling method merely samples the input every Δ seconds.

Once the generalized likelihood ratio is determined, we need to determine the threshold. If the *a priori* probabilities π_0 and π_1 are known, the evaluation of the threshold proceeds in the usual way. If they are not known, all of the conditional densities must not depend on the unknown parameters lest the performance probabilities also depend upon them. In most cases, the original model evaluation problem is posed in such a way that one of the models does not depend on the unknown parameter; a criterion on the performance probability related to that model can then be established via the Neyman-Pearson procedure. If not the case, the threshold cannot be computed and the threshold must be set experimentally: we force one of the models to be true and modify the threshold on the sufficient statistic until the desired level of performance is reached. Despite this non-mathematical approach, the overall performance of the model evaluation procedure will be optimum because of the results surrounding the Neyman-Pearson criterion.

5.4 Detection of Signals in Gaussian Noise

For the moment, we assume we know the joint distribution of the noise values. In most cases, the various models for the form of the observations—the hypotheses—do not differ because of noise characteristics. Rather, the signal component determines model variations and the noise is statistically independent of the signal; such is the specificity of detection problems in contrast to the generality of model evaluation. For example, we may want to determine whether a signal characteristic of a particular ship is present in a sonar array's output (the signal is known) or whether no ship is present (zero-valued signal).

To apply optimal hypothesis testing procedures previously derived, we first obtain a finite number L of observations—$X(l)$, $l = 0, \ldots, L-1$. These observations are usually obtained from continuous-time observations in one of two ways. Two commonly used methods for passing from continuous-time to discrete-time are known: *integrate-and-dump* and *sampling*. These techniques are illustrated in figure 5.8.

Integrate-and-Dump

In this procedure, no attention is paid to the bandwidth of the noise in selecting the sampling rate. Instead, the sampling interval Δ is selected according to the characteristics of the signal set. Because of the finite duration of the integrator, successive samples are statistically independent when the noise bandwidth exceeds $1/\Delta$. Consequently, the sampling rate can be varied to some extent while retaining this desirable analytic property.

Sampling

Traditional engineering considerations governed the selection of the sampling filter and the sampling

rate. As in the integrate-and-dump procedure, the sampling rate is chosen according to signal properties. Presumably, changes in sampling rate would force changes in the filter. As we shall see, this linkage has dramatic implications on performance.

With either method, the continuous-time detection problem of selecting between models (a binary selection is used here for simplicity)

$$\begin{aligned} \mathscr{M}_0: &\quad X(t) = s^0(t) + N(t) \quad 0 \le t < T \\ \mathscr{M}_1: &\quad X(t) = s^1(t) + N(t) \quad 0 \le t < T \end{aligned}$$

where $\{s_i(t)\}$ denotes the known signal set and $N(t)$ denotes additive noise modeled as a stationary stochastic process* is converted into the discrete-time detection problem

$$\begin{aligned} \mathscr{M}_0: &\quad X_l = s_l^0 + N_l \quad 0 \le l < L \\ \mathscr{M}_1: &\quad X_l = s_l^1 + N_l \quad 0 \le l < L \end{aligned}$$

where the sampling interval is *always* taken to divide the observation interval T: $L = T/\Delta$. We form the discrete-time observations into a vector: $\mathbf{X} = \text{col}[X(0), \ldots, X(L-1)]$. The binary detection problem is to distinguish between two possible signals present in the noisy output waveform.

$$\begin{aligned} \mathscr{M}_0: &\ \mathbf{X} = \mathbf{s}_0 + \mathbf{N} \\ \mathscr{M}_1: &\ \mathbf{X} = \mathbf{s}_1 + \mathbf{N} \end{aligned}$$

To apply our model evaluation results, we need the probability density of $\mathbf{X}$ under each model. As the only probabilistic component of the observations is the noise, the required density for the detection problem is given by

$$\boxed{p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i) = p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_i)}$$

and the corresponding likelihood ratio by

$$\boxed{\Lambda(\mathbf{X}) = \frac{p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_1)}{p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_0)}}.$$

Much of detection theory revolves about interpreting this likelihood ratio and deriving the detection threshold (either *threshold* or γ).

5.4.1 White Gaussian Noise

By far the easiest detection problem to solve occurs when the noise vector consists of statistically independent, identically distributed, Gaussian random variables. In this book, a "white" sequence consists of statistically independent random variables. The white sequence's mean is usually taken to be zero† and each component's variance is σ^2. The equal-variance assumption implies the noise characteristics are unchanging throughout the entire set of observations. The probability density of the zero-mean noise vector evaluated at $\mathbf{X} - \mathbf{s}_i$ equals that of Gaussian random vector having independent components ($\mathbf{K} = \sigma^2\mathbf{I}$) with mean $\mathbf{s}_i$.

$$p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{L/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{X} - \mathbf{s}_i)^t(\mathbf{X} - \mathbf{s}_i)\right\}$$

The resulting detection problem is similar to the Gaussian example examined so frequently in the hypothesis testing sections, with the distinction here being a non-zero mean under both models. The logarithm of the likelihood ratio becomes

$$(\mathbf{X} - \mathbf{s}_0)^t(\mathbf{X} - \mathbf{s}_0) - (\mathbf{X} - \mathbf{s}_1)^t(\mathbf{X} - \mathbf{s}_1) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} 2\sigma^2 \ln \eta$$

*We are *not* assuming the amplitude distribution of the noise to be Gaussian.

†The zero-mean assumption is realistic for the detection problem. If the mean were non-zero, simply subtracting it from the observed sequence results in a zero-mean noise component.

and the usual simplifications yield in

$$\left(\mathbf{X}^t\mathbf{s}_1 - \frac{\mathbf{s}_1^t\mathbf{s}_1}{2}\right) - \left(\mathbf{X}^t\mathbf{s}_0 - \frac{\mathbf{s}_0^t\mathbf{s}_0}{2}\right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta\,.$$

The quantities in parentheses express the signal processing operations for each model. If more than two signals were assumed possible, quantities such as these would need to be computed for each signal and the largest selected. This decision rule is optimum for the additive, white Gaussian noise problem.

Each term in the computations for the optimum detector has a signal processing interpretation. When expanded, the term $\mathbf{s}_i^t\mathbf{s}_i$ equals $\sum_{l=0}^{L-1} s_i^2(l)$, which is the *signal energy* E_i. The remaining term—$\mathbf{X}^t\mathbf{s}_i$—is the only one involving the observations and hence constitutes the sufficient statistic $\Upsilon_i(\mathbf{X})$ for the additive white Gaussian noise detection problem.

$$\boxed{\Upsilon_i(\mathbf{X}) = \mathbf{X}^t\mathbf{s}_i}$$

An abstract, but physically relevant, interpretation of this important quantity comes from the theory of linear vector spaces. There, the quantity $\mathbf{X}^t\mathbf{s}_i$ would be termed the *dot product* between $\mathbf{X}$ and $\mathbf{s}_i$ or the *projection* of $\mathbf{X}$ onto $\mathbf{s}_i$. By employing the Schwarz inequality, the largest value of this quantity occurs when these vectors are proportional to each other. Thus, a dot product computation measures how much alike two vectors are: they are completely alike when they are parallel (proportional) and completely dissimilar when orthogonal (the dot product is zero). More precisely, the dot product removes those components from the observations which are orthogonal to the signal. The dot product thereby generalizes the familiar notion of filtering a signal contaminated by broadband noise. In filtering, the signal-to-noise ratio of a bandlimited signal can be drastically improved by lowpass filtering; the output would consist only of the signal and "in-band" noise. The dot product serves a similar role, ideally removing those "out-of-band" components (the orthogonal ones) and retaining the "in-band" ones (those parallel to the signal).

Expanding the dot product, $\mathbf{X}^t\mathbf{s}_i = \sum_{l=0}^{L-1} X(l)s_i(l)$, another signal processing interpretation emerges. The dot product now describes a finite impulse response (FIR) filtering operation evaluated at a specific index. To demonstrate this interpretation, let $h(l)$ be the unit-sample response of a linear, shift-invariant filter where $h(l) = 0$ for $l < 0$ and $l \geq L$. Letting $X(l)$ be the filter's input sequence, the convolution sum expresses the output.

$$X(k) \otimes h(k) = \sum_{l=k-(L-1)}^{k} X(l)h(k-l)\,,$$

Letting $k = L-1$, the index at which the unit-sample response's last value overlaps the input's value at the origin, we have

$$X(k) \otimes h(k)\big|_{k=L-1} = \sum_{l=0}^{L-1} X(l)h(L-1-l).$$

If we set the unit-sample response equal to the index-reversed, then delayed signal $\big(h(l) = s_i(L-1-l)\big)$, we have

$$X(k) \otimes s_i(L-1-k)\big|_{k=L-1} = \sum_{l=0}^{L-1} X(l)s_i(l),$$

which equals the observation-dependent component of the optimal detector's sufficient statistic. Fig. 5.9 depicts these computations graphically. The sufficient statistic for the i^{th} signal is thus expressed in signal processing notation as $X(k) \otimes s_i(L-1-k)\big|_{k=L-1} - E_i/2$. The filtering term is called a *matched filter* because the observations are passed through a filter whose unit-sample response "matches" that of the signal being sought. We sample the matched filter's output at the precise moment when all of the observations fall within the filter's memory and then adjust this value by half the signal energy. The adjusted values for the two assumed signals are subtracted and compared to a threshold.

To compute the performance probabilities, the expressions should be simplified in the ways discussed in the hypothesis testing sections. As the energy terms are known *a priori*, they can be incorporated into the

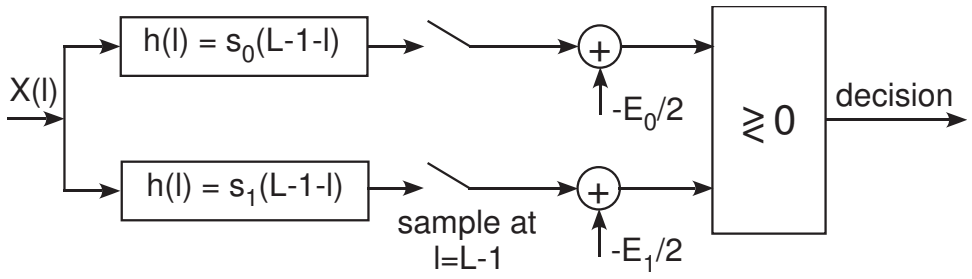


Figure 5.9: The detector for signals contained in additive, white Gaussian noise consists of a matched filter, whose output is sampled at the duration of the signal and half of the signal energy is subtracted from it. The optimum detector incorporates a matched filter for each signal compares their outputs to determine the largest.

threshold with the result

$$\sum_{l=0}^{L-1} X(l)[s_1(l)-s_0(l)] \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta + \frac{E_1-E_0}{2}.$$

The left term constitutes the sufficient statistic for the binary detection problem. Because the additive noise is presumed Gaussian, the sufficient statistic is a Gaussian random variable no matter which model is assumed. Under $\mathscr{M}_i$, the specifics of this probability distribution are

$$\sum_{l=0}^{L-1} X(l)[s_1(l)-s_0(l)] \sim \mathscr{N}\left(\sum s_i(l)[s_1(l)-s_0(l)], \sigma^2 \sum [s_1(l)-s_0(l)]^2\right).$$

The false-alarm probability is given by

$$P_F = Q\left(\frac{\sigma^2 \ln\eta + (E_1-E_0)/2 - \sum s_0(l)[s_1(l)-s_0(l)]}{\sigma \cdot \{\sum [s_1(l)-s_0(l)]^2\}^{1/2}}\right).$$

The signal-related terms in the numerator of this expression can be manipulated with the false-alarm probability (and the detection probability) for the optimal white Gaussian noise detector succinctly expressed by

$$\boxed{\begin{aligned} P_F &= Q\left(\frac{\ln\eta + \frac{1}{2\sigma^2}\sum [s_1(l)-s_0(l)]^2}{\frac{1}{\sigma}\{\sum [s_1(l)-s_0(l)]^2\}^{1/2}}\right) \\ P_D &= Q\left(\frac{\ln\eta - \frac{1}{2\sigma^2}\sum [s_1(l)-s_0(l)]^2}{\frac{1}{\sigma}\{\sum [s_1(l)-s_0(l)]^2\}^{1/2}}\right) \end{aligned}}$$

Note that the *only* signal-related quantity affecting this performance probability (and all of the others) is the *ratio of energy in the difference signal to the noise variance*. The larger this ratio, the better (smaller) the performance probabilities become. Note that the details of the signal waveforms do not greatly affect the energy of the difference signal. For example, consider the case where the two signal energies are equal ($E_0 = E_1 = E$); the energy of the difference signal is given by $2E - 2\sum s_0(l)s_1(l)$. The largest value of this energy occurs when the signals are negatives of each other, with the difference-signal energy equaling $4E$. Thus, equal-energy but opposite-signed signals such as sine waves, square-waves, Bessel functions, etc. *all* yield exactly the same performance levels. The essential signal properties that do yield good performance values are elucidated by an alternate interpretation. The term $\sum [s_1(l)-s_0(l)]^2$ equals $\|\mathbf{s}_1 - \mathbf{s}_0\|^2$, the L^2 norm of the difference signal. Geometrically, the difference-signal energy is the same quantity as the square of the Euclidean distance between the two signals. In these terms, a larger distance between the two signals will mean better performance.

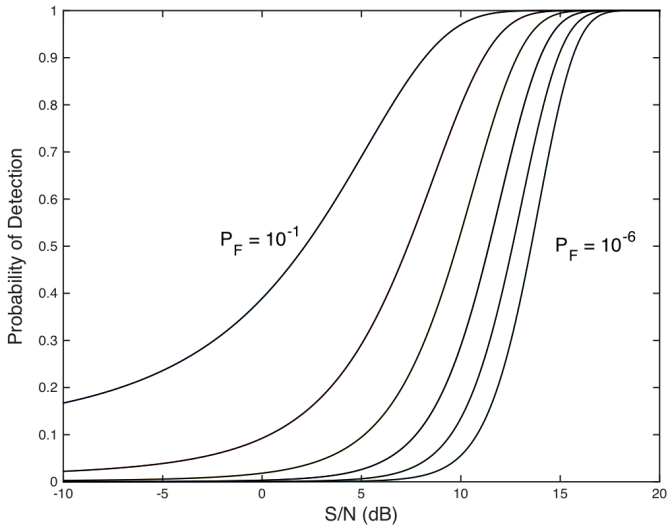


Figure 5.10: The probability of detection is plotted versus signal-to-noise ratio for various values of the false-alarm probability P_F. False-alarm probabilities range from 10^{-1} down to 10^{-6} by decades. The matched filter receiver was used since the noise is white and Gaussian. Note how the range of signal-to-noise ratios over which the detection probability changes shrinks as the false-alarm probability decreases. This effect is a consequence of the non-linear nature of the function $Q(\cdot)$.

Example

A common detection problem is to determine whether a signal is present ($\mathscr{M}_1$) or not ($\mathscr{M}_0$) in an observed signal. In this case, $s_0(l) = 0$. The optimal detector relies on filtering the observations with a matched filter having an impulse response based on the assumed signal. Letting the signal under $\mathscr{M}_1$ be denoted simply by $s(l)$, the optimal detector consists of

$$X(l) \otimes s(L-1-l)|_{l=L-1} - E/2 \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln \eta$$

$$\text{or} \quad X(l) \otimes s(L-1-l)|_{l=L-1} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma.$$

The false-alarm and detection probabilities are given by

$$P_F = Q\left(\frac{\gamma}{E^{1/2}/\sigma}\right) \quad P_D = Q\left(Q^{-1}(P_F) - \sqrt{\frac{E}{\sigma}}\right).$$

Fig. 5.10 displays the probability of detection as a function of the signal-to-noise ratio E/σ^2 for several values of false-alarm probability. Given an estimate of the expected signal-to-noise ratio, these curves can be used to assess the trade-off between the false-alarm and detection probabilities.

The important parameter determining detector performance derived in this example is the *signal-to-noise ratio* E/σ^2: the larger it is, the smaller the false-alarm probability is (generally speaking). Signal-to-noise ratios can be measured in many different ways. For example, one measure might be the ratio of the rms signal amplitude to the rms noise amplitude. Note that the important one for the detection problem is much different. The signal portion is the *sum* of the squared signal values over the *entire* set of observed values—the signal energy; the noise portion is the variance of *each* noise component—the noise power. Thus, energy

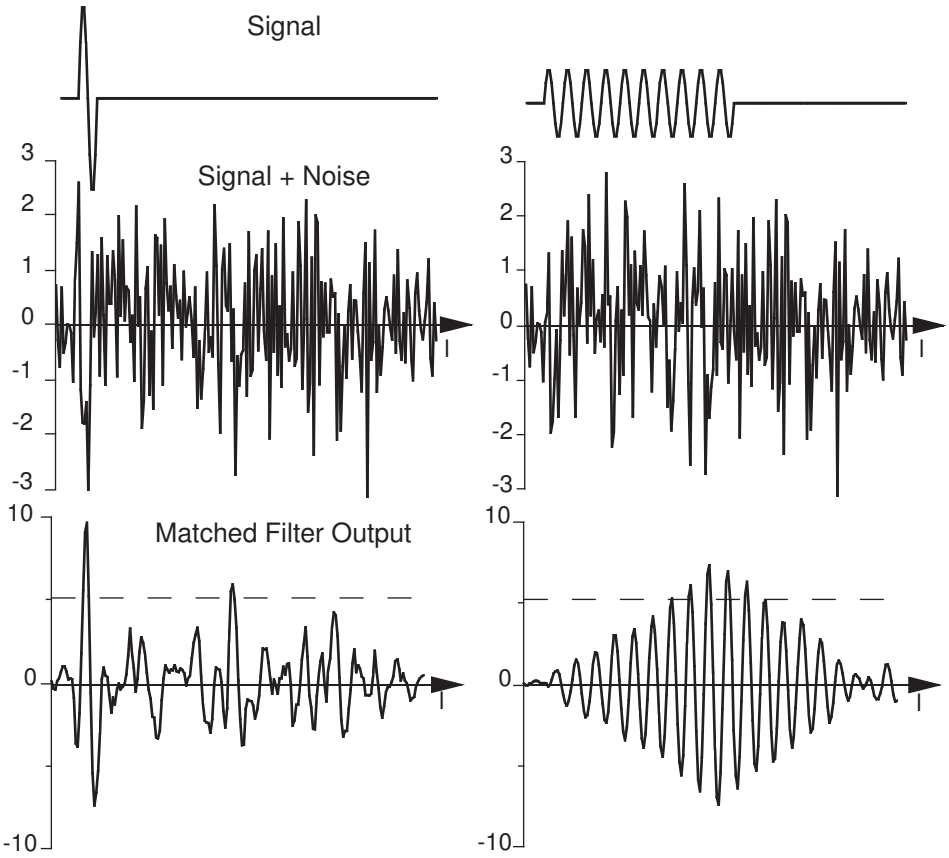


Figure 5.11: Two signals having the same energy are shown at the top of the figure. The one on the left equals one cycle of a sinusoid having ten samples/period ($\sin(2\pi f_o l)$ with $f_o = 0.1$). On the right, ten cycles of similar signal is shown, with an amplitude a factor of $\sqrt{10}$ smaller. The middle portion of the figure shows these signals with the same noise signal added; the duration of this signal is 200 samples. The lower portion depicts the outputs of matched filters for each signal. The detection threshold was set by specifying a false-alarm probability of 10^{-2}.

can be increased in two ways that increase the signal-to-noise ratio: the signal can be made larger *or* the observations can be extended to encompass a larger number of values.

To illustrate this point, two signals having the same energy are shown in Fig. 5.11. When these signals are shown in the presence of additive noise, the signal is visible on the left because its amplitude is larger; the one on the right is much more difficult to discern. The instantaneous signal-to-noise ratio—the ratio of signal amplitude to average noise amplitude—is the important visual cue. However, the kind of signal-to-noise ratio that determines detection performance belies the eye. The matched filter outputs have similar maximal values, indicating that total signal energy rather than amplitude determines the performance of a matched filter detector.

Validity of the White Noise Assumption

The optimal detection paradigm for the additive, white Gaussian noise problem has a relatively simple solution: construct FIR filters whose unit-sample responses are related to the presumed signals and compare the filtered outputs with a threshold. We may well wonder which assumptions made in this problem are most questionable in "real-world" applications. Noise is additive in most cases. In many situations, the additive noise present in observed data is Gaussian. Because of the Central Limit Theorem, if numerous noise sources

impinge on a measuring device, their superposition will be Gaussian to a great extent. As we know from the discussion in §2.1.10 {11}, glibly appealing to the Central Limit Theorem is not without hazards; the non-Gaussian detection problem will be discussed in some detail later. Interestingly, the weakest assumption is the "whiteness" of the noise. Note that the observation sequence is obtained as a result of *sampling* the sensor outputs. Assuming white noise samples does *not* mean that the continuous-time noise was white. White noise in continuous time has infinite variance and cannot be sampled; discrete-time white noise has a finite variance with a constant power spectrum. The Sampling Theorem suggests that a signal is represented accurately by its samples only if we choose a sampling frequency commensurate with the signal's bandwidth. One should note that fidelity of representation does *not* mean that the sample values are independent. In most cases, satisfying the Sampling Theorem means that the samples are correlated. As shown in §2.2.3 {14}, the correlation function of sampled noise equals samples of the original correlation function. For the sampled noise to be white, $\mathsf{E}[N(l_1T)N(l_2T)] = 0$ for $l_1 \neq l_2$: the samples of the correlation function at locations other than the origin must all be zero. While some correlation functions have this property, *many examples satisfy the sampling theorem but do not yield uncorrelated samples.* In many practical situations, *undersampling* the noise will reduce inter-sample correlation. Thus, we obtain uncorrelated samples either by deliberately undersampling, which wastes signal energy, or by imposing anti-aliasing filters that have a bandwidth larger than the signal and sampling at the signal's Nyquist rate. Since the noise power spectrum usually extends to higher frequencies than the signal, this intentional undersampling can result in larger noise variance. In either case, by trying to make the problem at hand match the solution, we are actually reducing performance! We need a *direct* approach to attacking the correlated noise issue that arises in virtually *all* sampled-data detection problems rather than trying to work around it.

5.4.2 Colored Gaussian Noise

When the additive Gaussian noise in the sensors' outputs is colored (i.e., the noise values are correlated in some fashion), the linearity of beamforming algorithms means that the array processing output X also contains colored noise. The solution to the colored-noise, binary detection problem remains the likelihood ratio, but differs in the form of the *a priori* densities. The noise will again be assumed zero mean, but the noise vector has a non-trivial covariance matrix $\mathbf{K}$: $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$.

$$p_{\mathbf{N}}(\mathbf{N}) = \frac{1}{\sqrt{\det[2\pi\mathbf{K}]}} \exp\left\{-\frac{1}{2}\mathbf{N}^t\mathbf{K}^{-1}\mathbf{N}\right\}$$

In this case, the logarithm of the likelihood ratio is

$$(\mathbf{X}-\mathbf{s}_1)^t\mathbf{K}^{-1}(\mathbf{X}-\mathbf{s}_1) - (\mathbf{X}-\mathbf{s}_0)^t\mathbf{K}^{-1}(\mathbf{X}-\mathbf{s}_0) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} 2\ln\eta$$

which, after the usual simplifications, is written

$$\left[\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_1 - \frac{\mathbf{s}_1^t\mathbf{K}^{-1}\mathbf{s}_1}{2}\right] - \left[\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_0 - \frac{\mathbf{s}_0^t\mathbf{K}^{-1}\mathbf{s}_0}{2}\right] \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta\,.$$

The sufficient statistic for the colored Gaussian noise detection problem is

$$\boxed{\Upsilon_i(\mathbf{X}) = \mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i}. \tag{5.13}$$

The quantities computed for each signal have a similar, but more complicated interpretation than in the white noise case. $\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i$ is a dot product, but with respect to the so-called *kernel* $\mathbf{K}^{-1}$. The effect of the kernel is to weight certain components more heavily than others. A positive-definite symmetric matrix (the covariance matrix is one such example) can be expressed in terms of its eigenvectors and eigenvalues.

$$\mathbf{K}^{-1} = \sum_{k=1}^{L} \frac{1}{\lambda_k}\mathbf{v}_k\mathbf{v}_k^t$$

The sufficient statistic can thus be written as the complicated summation

$$\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i = \sum_{k=1}^{L} \frac{1}{\lambda_k}(\mathbf{X}^t\mathbf{v}_k)(\mathbf{v}_k^t\mathbf{s}_i)\,,$$

where λ_k and $\mathbf{v}_k$ denote the k^{th} eigenvalue and eigenvector of the covariance matrix $\mathbf{K}$. Each of the constituent dot products is largest when the signal and the observation vectors have strong components parallel to $\mathbf{v}_k$. However, the product of these dot products is weighted by the reciprocal of the associated eigenvalue. Thus, components in the observation vector parallel to the signal will tend to be accentuated; those components parallel to the eigenvectors having the *smaller* eigenvalues will receive greater accentuation than others. The usual notions of parallelism and orthogonality become "skewed" because of the presence of the kernel. A covariance matrix's eigenvalue has "units" of variance; these accentuated directions thus correspond to small noise variances. We can therefore view the weighted dot product as a computation that is simultaneously trying to select components in the observations similar to the signal, but concentrating on those where the noise variance is small.

The second term in the expressions constituting the optimal detector are of the form $\mathbf{s}_i^t\mathbf{K}^{-1}\mathbf{s}_i$. This quantity is a special case of the dot product just discussed. The two vectors involved in this dot product are identical; they are parallel by definition. The weighting of the signal components by the reciprocal eigenvalues remains. Recalling the units of the eigenvectors of $\mathbf{K}$, $\mathbf{s}_i^t\mathbf{K}^{-1}\mathbf{s}_i$ has the units of a signal-to-noise ratio, which is computed in a way that enhances the contribution of those signal components parallel to the "low noise" directions.

To compute the performance probabilities, we express the detection rule in terms of the sufficient statistic.

$$\mathbf{X}^t\mathbf{K}^{-1}(\mathbf{s}_1-\mathbf{s}_0) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta + \frac{1}{2}(\mathbf{s}_1^t\mathbf{K}^{-1}\mathbf{s}_1 - \mathbf{s}_0^t\mathbf{K}^{-1}\mathbf{s}_0)$$

The distribution of the sufficient statistic on the left side of this equation is Gaussian because it consists as a linear transformation of the Gaussian random vector $\mathbf{X}$. Assuming the i^{th} model to be true,

$$\mathbf{X}^t\mathbf{K}^{-1}(\mathbf{s}_1-\mathbf{s}_0) \sim \mathscr{N}\left(\mathbf{s}_i^t\mathbf{K}^{-1}(\mathbf{s}_1-\mathbf{s}_0), (\mathbf{s}_1-\mathbf{s}_0)^t\mathbf{K}^{-1}(\mathbf{s}_1-\mathbf{s}_0)\right).$$

The false-alarm probability for the optimal Gaussian colored noise detector is given by

$$\boxed{P_F = Q\left(\frac{\ln\eta + \frac{1}{2}(\mathbf{s}_1-\mathbf{s}_0)^t\mathbf{K}^{-1}(\mathbf{s}_1-\mathbf{s}_0)}{[(\mathbf{s}_1-\mathbf{s}_0)^t\mathbf{K}^{-1}(\mathbf{s}_1-\mathbf{s}_0)]^{1/2}}\right)}. \tag{5.14}$$

As in the white noise case, the important signal-related quantity in this expression is the signal-to-noise ratio of the difference signal. The distance interpretation of this quantity remains, but the distance is now warped by the kernel's presence in the dot product.

The sufficient statistic computed for each signal can be given two signal processing interpretations in the colored noise case. Both of these rest on considering the quantity $\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i$ as a simple dot product, but with different ideas on grouping terms. The simplest is to group the kernel with the signal so that the sufficient statistic is the dot product between the observations and a *modified* version of the signal $\tilde{\mathbf{s}}_i = \mathbf{K}^{-1}\mathbf{s}_i$. This modified signal thus becomes the equivalent to the unit-sample response of the matched filter. In this form, the observed data are unaltered and passed through a matched filter whose unit-sample response depends on both the signal and the noise characteristics. The size of the noise covariance matrix, equal to the number of observations used by the detector, is usually large: hundreds if not thousands of samples are possible. Thus, computation of the inverse of the noise covariance matrix becomes an issue. This problem needs to be solved only once if the noise characteristics are static; the inverse can be precomputed on a general purpose computer using well-established numerical algorithms. The signal-to-noise ratio term of the sufficient statistic is the dot product of the signal with the modified signal $\tilde{\mathbf{s}}_i$. This view of the receiver structure is shown in Fig. 5.12.

A second and more theoretically powerful view of the computations involved in the colored noise detector emerges when we *factor* covariance matrix. The *Cholesky factorization* of a positive-definite, symmetric

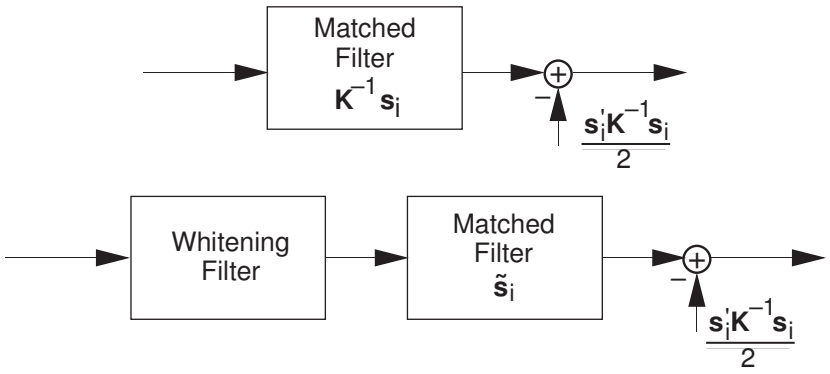


Figure 5.12: These diagrams depict the signal processing operations involved in the optimum detector when the additive noise is not white. The upper diagram shows a matched filter whose unit-sample response depends both on the signal and the noise characteristics. The lower diagram is often termed the whitening filter structure, where the noise components of the observed data are first whitened, then passed through a matched filter whose unit-sample response is related to the "whitened" signal.

matrix (such as a covariance matrix or its inverse) has the form $\mathbf{K} = \mathbf{LDL}^t$. With this factorization, the sufficient statistic can be written as

$$\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i = \left(\mathbf{D}^{-1/2}\mathbf{L}^{-1}\mathbf{X}\right)^t\left(\mathbf{D}^{-1/2}\mathbf{L}^{-1}\mathbf{s}_i\right).$$

The components of the dot product are multiplied by the same matrix ($\mathbf{D}^{-1/2}\mathbf{L}^{-1}$), which is lower-triangular. *If* this matrix were also Toeplitz, the product of this kind between a Toeplitz matrix and a vector would be equivalent to the convolution of the components of the vector with the first column of the matrix. If the matrix is not Toeplitz (which, inconveniently, is the typical case), a convolution also results, but with a unit-sample response that varies with the index of the output—a time-varying, linear filtering operation. The variation of the unit-sample response corresponds to the different rows of the matrix $\mathbf{D}^{-1/2}\mathbf{L}^{-1}$ running *backwards* from the main-diagonal entry. What is the physical interpretation of the action of this filter? The covariance of the random vector $\mathbf{x} = \mathbf{AX}$ is given by $\mathbf{K}_x = \mathbf{AK}_X\mathbf{A}^t$. Applying this result to the current situation, we set $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{L}^{-1}$ and $\mathbf{K}_X = \mathbf{K} = \mathbf{LDL}^t$ with the result that the covariance matrix $\mathbf{K}_x$ is the identity matrix! Thus, the matrix $\mathbf{D}^{-1/2}\mathbf{L}^{-1}$ corresponds to a (possibly time-varying) *whitening filter*: we have converted the colored-noise component of the observed data to white noise! As the filter is always linear, the Gaussian observation noise remains Gaussian at the output. Thus, the colored noise problem is converted into a simpler one with the whitening filter: the whitened observations are first match-filtered with the "whitened" signal $\mathbf{s}_i^+ = \mathbf{D}^{-1/2}\mathbf{L}^{-1}\mathbf{s}_i$ (whitened with respect to noise characteristics only) then half the energy of the whitened signal is subtracted (Fig. 5.12).

Example

To demonstrate the interpretation of the Cholesky factorization of the covariance matrix as a time-varying whitening filter, consider the covariance matrix

$$\mathbf{K} = \begin{bmatrix} 1 & a & a^2 & a^3 \\ a & 1 & a & a^2 \\ a^2 & a & 1 & a \\ a^3 & a^2 & a & 1 \end{bmatrix}.$$

This covariance matrix indicates that the noise was produced by passing white Gaussian noise through a first-order filter having coefficient a: $N(l) = aN(l-1) + \left(1-a^2\right)^{1/2}w(l)$, where $w(l)$ is unit-variance white noise. Thus, we would expect that if a whitening filter emerged from the matrix manipulations (derived just below), it would be a first-order FIR filter having an unit-sample response

proportional to

$$h(l) = \begin{cases} 1, & l = 0 \\ -a, & l = 1 \\ 0, & \text{otherwise}. \end{cases}$$

Simple arithmetic calculations of the Cholesky decomposition suffice to show that the matrices $\mathbf{L}$ and $\mathbf{D}$ are given by

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ a & 1 & 0 & 0 \\ a^2 & a & 1 & 0 \\ a^3 & a^2 & a & 1 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-a^2 & 0 & 0 \\ 0 & 0 & 1-a^2 & 0 \\ 0 & 0 & 0 & 1-a^2 \end{bmatrix}$$

and that their inverses are

$$\mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -a & 1 & 0 & 0 \\ 0 & -a & 1 & 0 \\ 0 & 0 & -a & 1 \end{bmatrix} \quad \mathbf{D}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{1-a^2} & 0 & 0 \\ 0 & 0 & \frac{1}{1-a^2} & 0 \\ 0 & 0 & 0 & \frac{1}{1-a^2} \end{bmatrix}.$$

Because $\mathbf{D}$ is diagonal, the matrix $\mathbf{D}^{-1/2}$ equals the term-by-term square root of the inverse of $\mathbf{D}$. The product of interest here is therefore given by

$$\mathbf{D}^{-1/2}\mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{-a}{\sqrt{1-a^2}} & \frac{1}{\sqrt{1-a^2}} & 0 & 0 \\ 0 & \frac{-a}{\sqrt{1-a^2}} & \frac{1}{\sqrt{1-a^2}} & 0 \\ 0 & 0 & \frac{-a}{\sqrt{1-a^2}} & \frac{1}{\sqrt{1-a^2}} \end{bmatrix}.$$

Let $\tilde{\mathbf{X}}$ express the product $\mathbf{D}^{-1/2}\mathbf{L}^{-1}\mathbf{X}$. This vector's elements are given by

$$\tilde{X}_0 = X_0, \quad \tilde{X}_1 = \frac{1}{\sqrt{1-a^2}}[X_1 - aX_0], \quad etc.$$

Thus, the expected FIR whitening filter emerges after the first term. The expression could *not* be of this form as no observations were assumed to precede X_0. This edge effect is the source of the time-varying aspect of the whitening filter. If the system modeling the noise generation process has only poles, this whitening filter will always stabilize—not vary with time—once sufficient data are present within the memory of the FIR inverse filter. In contrast, the presence of zeros in the generation system would imply an IIR whitening filter. With finite data, the unit-sample response would then change on each output sample.

5.4.3 Spectral Detection

From the results presented in the previous sections, the colored noise problem was found to be pervasive, but required a computationally difficult detector. The simplest detector structure occurs when the additive noise is white; this notion lead to the idea of whitening the observations, thereby transforming the data into a simpler form (as far as detection theory is concerned). However, the required whitening filter is often time-varying and can have a long-duration unit-sample response. Other, more computationally expedient, approaches to whitening are worth considering. An only slightly more complicated detection problem occurs when we have a diagonal noise covariance matrix, as in the white noise case, but unequal values on the diagonal. In terms of the observations, this situation means that they are contaminated by noise having statistically independent,

but unequal variance components: the noise would thus be non-stationary. Few problems fall directly into this category; however, the colored noise problem can be recast into the white, unequal-variance problem by calculating the discrete Fourier Transform (DFT) of the observations and basing the detector on the resulting spectrum. The resulting *spectral detectors* greatly simplify detector structures for discrete-time problems *if* the qualifying assumptions described in sequel hold.

Let $\mathbf{W}$ be the so-called $L \times L$ "DFT matrix"

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W & W^2 & \cdots & W^{L-1} \\ 1 & W^2 & W^4 & \cdots & W^{2(L-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & W^{L-1} & W^{2(L-1)} & \cdots & W^{(L-1)(L-1)} \end{bmatrix}$$

where W is the elementary complex exponential $\exp\{-j\frac{2\pi}{L}\}$. The discrete Fourier Transform of the sequence $x(l)$, usually written as $X(k) = \sum_{l=0}^{L-1} x(l)\exp\{-j\frac{2\pi lk}{L}\}$, can be written in matrix form as $\mathbf{X} = \mathbf{W}\mathbf{x}$. To analyze the effect of evaluating the DFT of the observations, we describe the computations in matrix form for analytic simplicity. The first critical assumption has been made: take special note that the length of the transform *equals* the duration of the observations. In many signal processing applications, the transform length can differ from the data length, being either longer or shorter. The statistical properties developed in the following discussion are critically sensitive to the equality of these lengths. The covariance matrix $\mathbf{K}_X$ of $\mathbf{X}$ is given by $\mathbf{W}\mathbf{K}_x\mathbf{W}'$. Symmetries of these matrices—the Vandermonde form of $\mathbf{W}$ and the Hermitian, Toeplitz form of $\mathbf{K}_x$—leads to many simplifications in evaluating this product. The entries on the main diagonal are given by*

$$K^X_{kk} = \sum_{l=-(L-1)}^{(L-1)} (L-|l|)K^x_{1,|l|+1}e^{-j\frac{2\pi lk}{L}}.$$

The *variance* of the k^{th} term in the discrete Fourier Transform of the noise thus equals the discrete Fourier Transform of the *windowed* covariance function. This window has a triangular shape; colloquially termed the "rooftop" window, its technical name is the Bartlett window and it occurs frequently in spectral estimation (see page 115). We have found that the variance equals the smoothed noise power spectrum evaluated at a particular frequency. The off-diagonal terms of $\mathbf{K}_X$ are not as easily written; the complicated result is

$$K^X_{k_1k_2} = \sum_{l=0}^{L-1} K^x_{1,l+1} \frac{(-1)^{k_1-k_2+1}\sin\frac{\pi l(k_1-k_2)}{L}}{\sin\frac{\pi(k_1-k_2)}{L}} \left(e^{+j\frac{2\pi lk_1}{L}} + e^{-j\frac{2\pi lk_2}{L}}\right), \quad k_1 \neq k_2\,.$$

The complex exponential terms indicate that each off-diagonal term consists of the sum of two Fourier Transforms: one at the frequency index k_2 and the other negative index $-k_1$. In addition, the transform is evaluated only over non-negative lags. The transformed quantity again equals a windowed version of the noise covariance function, but with a *sinusoidal* window whose frequency depends on the indices k_1 and k_2. This window can be negative-valued! In contrast to the Bartlett window encountered in evaluating the on-diagonal terms, the maximum value achieved by the window is not large ($1/\sin\frac{\pi(k_1-k_2)}{L}$ compared to L). Furthermore, this window is *always* zero at the origin, the location of the maximum value of any covariance function. The largest magnitudes of the off-diagonal terms tend to occur when the indices k_1 and k_2 are nearly equal. Let their difference be one; if the covariance function of the noise tends toward zero well within the number of observations L, then the Bartlett window has little effect on the covariance function while the sinusoidal window greatly reduces it. This condition on the covariance function can be interpreted physically: the noise in this case is wideband and any correlation between noise values does not extend over significant portion of the observation record. On the other hand, if the width of the the covariance function is comparable to

*The curious index $l+1$ on the matrix arises because rows and columns of matrices are traditionally indexed beginning with one instead of zero.

L, the off-diagonal terms will be significant. This situation occurs when the noise bandwidth is smaller than or comparable to the reciprocal of the observation interval's duration. This condition on the duration of the observation interval relative to the width of the noise correlation function forms the second critical assumption of spectral detection. The off-diagonal terms will thus be much smaller than corresponding terms on the main diagonal ($\left|K^X_{k_1k_2}\right|^2 \ll K^X_{k_1k_1}K^X_{k_2k_2}$).

In the simplest case, the covariance matrix of the discrete Fourier Transform of the observations can be well approximated by a diagonal matrix.

$$\mathbf{K}_X = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{L-1}^2 \end{bmatrix}$$

The non-zero components σ_k^2 of this matrix constitute the noise power spectrum at the various frequencies. The signal component of the transformed observations $\mathbf{X}$ is represented by $\mathbf{S}_i$, the DFT of the signal $\mathbf{s}_i$, while the noise component has this diagonal covariance matrix structure. *Thus, in the frequency domain, the colored noise problem can be approximately converted to a white noise problem where the components of the noise have unequal variances*. To recap, the critical assumptions of spectral detection are

- The transform length equals that of the observations. In particular, the observations cannot be "padded" to force the transform length to equal a "nice" number (like a power of two).
- The noise's correlation structure should be much less than the duration of the observations. Equivalently, a narrow correlation function means the corresponding power spectrum varies slowly with frequency. If either condition fails to hold, calculating the Fourier Transform of the observations will not necessarily yield a simpler noise covariance matrix.

The optimum spectral detector computes, for each possible signal, the quantity $\mathrm{Re}[\mathbf{X}'\mathbf{K}_X^{-1}\mathbf{S}_i] - \mathbf{S}_i'\mathbf{K}_X^{-1}\mathbf{S}_i/2$.* Because of the covariance matrix's simple form, this sufficient statistic for the spectral detection problem has the simple form

$$\boxed{\mathrm{Re}[\mathbf{X}'\mathbf{K}_X^{-1}\mathbf{S}_i] - \frac{1}{2}\mathbf{S}_i'\mathbf{K}_X^{-1}\mathbf{S}_i = \sum_{k=0}^{L-1}\left(\frac{\mathrm{Re}[X^*(k)S_i(k)]}{\sigma_k^2} - \frac{1}{2}\frac{|S_i(k)|^2}{\sigma_k^2}\right)}. \tag{5.15}$$

Each term in the dot product between the discrete Fourier Transform of the observations and the signal is weighted by the reciprocal of the noise power spectrum at that frequency. This computation is much simpler than the equivalent time domain version and, because of algorithms such as the fast Fourier Transform, the initial transformation (the multiplication by $\mathbf{W}$ or the discrete Fourier Transform) can be evaluated expeditiously.

Sinusoidal signals are particularly well-suited to the spectral detection approach. *If* the signal's frequency equals one of the analysis frequencies in the Fourier Transform ($f_o = k/L$ for some k), then the sequence $S_i(k)$ is non-zero only at this frequency index, only one term in the sufficient statistic's summation need be computed, and the noise power is no longer explicitly needed by the detector (it can be merged into the threshold).

$$\mathrm{Re}[\mathbf{X}'\mathbf{K}_X^{-1}\mathbf{S}_i] - \frac{1}{2}\mathbf{S}_i'\mathbf{K}_X^{-1}\mathbf{S}_i = \frac{\mathrm{Re}[X^*(k)S_i(k)]}{\sigma_k^2} - \frac{1}{2}\frac{|S_i(k)|^2}{\sigma_k^2}$$

If the signal's frequency does not correspond to one of the analysis frequencies, spectral energy will be maximal at the nearest analysis frequency but will extend to nearby frequencies also. This effect is termed "leakage" and has been well studied. Exact formulation of the signal's DFT is usually complicated in this case; approximations which utilize only the maximal-energy frequency component will be sub-optimal (i.e., yield a smaller detection probability). The performance reduction may be small, however, justifying the reduced amount of computation.

*The real part in the statistic emerges because $\mathbf{X}$ and $\mathbf{S}_i$ are complex quantities.

5.5 Detection in the Presence of Uncertainties

5.5.1 Unknown Signal Parameters

Applying the techniques described in the previous section may be difficult to justify when the signal and/or noise models are uncertain. For example, we must "know" a signal down to the precise value of every sample. In other cases, we may know the signal's waveform, but not the waveform's *amplitude* as measured by a sensor. A ubiquitous example of this uncertainty is propagation loss: the range of a far-field signal can only be lower-bounded, which leads to the known waveform, unknown amplitude detection problem. Another uncertainty is the signal's *time origin*: Without this information, we do not know when to start the matched filtering operation! In other circumstances, the noise may have a white power spectrum but its variance is unknown. Much worse situations (from the point of view of detection theory) can be imagined: the signal may not be known at all and one may want to detect the presence of *any* disturbance in the observations other than that of well-specified noise. These problems are very realistic, but the detection schemes as presented are inadequate to attack them. The detection results we have derived to date need to be extended to incorporate the presence of unknowns just as we did in hypothesis testing (§5.3 {169}).

Unknown Signal Amplitude

Assume that a signal's waveform is known exactly, but the amplitude is not. We need an algorithm to detect the presence or absence of this signal observed in additive noise at an array's output. The models can be formally stated as

$$\begin{array}{ll} \mathscr{M}_0: & X(l) = N(l) \\ \mathscr{M}_1: & X(l) = As(l) + N(l), \quad A = ? \end{array} \qquad l = 0, \ldots, L-1 .$$

As usual, L observations are available and the noise is Gaussian. This problem is equivalent to an unknown parameter problem described in §5.3 {169}. We learned there that the first step is to ascertain the existence of a uniformly most powerful test. For each value of the unknown parameter A, the logarithm of the likelihood ratio is written

$$A\,\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s} - \frac{A^2\mathbf{s}^t\mathbf{K}^{-1}\mathbf{s}}{2} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta$$

Assuming that $A > 0$, a typical assumption in array processing problems, we write this comparison as

$$\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{1}{A}\ln\eta + \frac{A\mathbf{s}^t\mathbf{K}^{-1}\mathbf{s}}{2} = \gamma$$

As the sufficient statistic does not depend on the unknown parameter and one of the models ($\mathscr{M}_0$) does not depend on this parameter, a uniformly most powerful test exists: the threshold term, despite its explicit dependence on a variety of factors, can be determined by specifying a false-alarm probability. If the noise is not white, the whitening filter or a spectral transformation may be used to simplify the computation of the sufficient statistic.

Example

Assume that the waveform, but not the amplitude of a signal is known. The Gaussian noise is white with a variance of σ^2. The decision rule expressed in terms of a sufficient statistic becomes

$$\mathbf{X}^t\mathbf{s} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma .$$

The false-alarm probability is given by

$$P_F = Q\left(\frac{\gamma}{\sqrt{E\sigma^2}}\right),$$

where E is the *assumed* signal energy which equals $\|\mathbf{s}\|^2$. The threshold γ is thus found to be

$$\gamma = \sqrt{E\sigma^2}\,Q^{-1}(P_F).$$

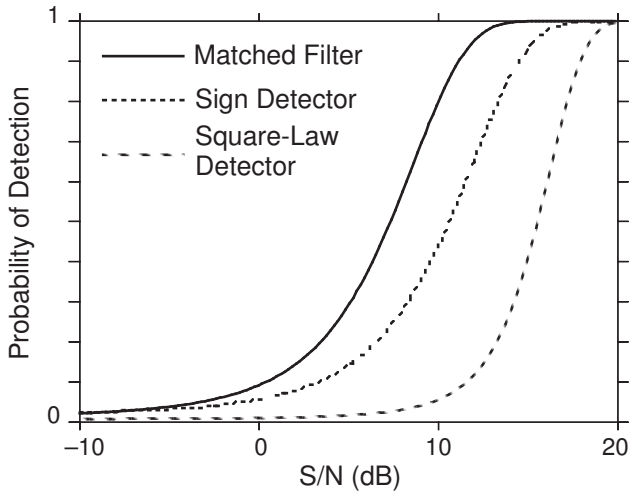


Figure 5.13: The false-alarm probability of the detector was fixed at 10^{-2}. The signal equaled $A\sin(2\pi f_o l)$, $l = 0,\ldots,L-1$, where f_o was 0.1 and $L = 100$; the noise was white and Gaussian. The detection probabilities that result from a matched filter detector, a sign detector, and a square-law detector are shown. These detectors make progressively fewer assumptions about the signal, consequently yielding progressively smaller detection probabilities.

The probability of detection for the matched filter detector is given by

$$\begin{aligned} P_D &= Q\left(\frac{\gamma - AE}{\sqrt{E\sigma^2}}\right) \\ &= Q\left(Q^{-1}(P_F) - \sqrt{\frac{A^2E}{\sigma^2}}\right), \end{aligned}$$

where A is the signal's *actual* amplitude relative to the assumed signal having energy E. Thus, the observed signal, when it is present, has energy A^2E. The probability of detection is shown in Fig. 5.13 as a function of the observed signal-to-noise ratio. For any false-alarm probability, the signal must be sufficiently energetic for its presence to be reliably determined.

All too many interesting problems exist where a uniformly most powerful decision rule cannot be found. Suppose in the problem just described that the amplitude is known ($A = 1$, for example), but the variance of the noise is not. Writing the covariance matrix as $\sigma^2\widetilde{\mathbf{K}}$, where we normalize the covariance matrix to have unit variance entries by requiring $\mathrm{tr}[\widetilde{\mathbf{K}}] = L$, unknown values of σ^2 express the known correlation structure, unknown noise power problem. From the results just given, the decision rule can be written so that the sufficient statistic does not depend on the unknown variance.

$$\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{s} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta + \frac{\mathbf{s}^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}}{2} = \gamma\,.$$

However, as *both* models depend on the unknown parameter, performance probabilities cannot be computed and we cannot design a detection threshold.

Hypothesis testing ideas show the way out; estimate the unknown parameter(s) under each model separately and then use these estimates in the likelihood ratio (Eq. 5.12 {173}). Using the maximum likelihood estimates for the parameters results in the generalized likelihood ratio test for the detection problem [61, 62, 89]. Letting $\boldsymbol{\theta}$ denote the vector of unknown parameters, be they for the signal or the noise, the generalized likeli-

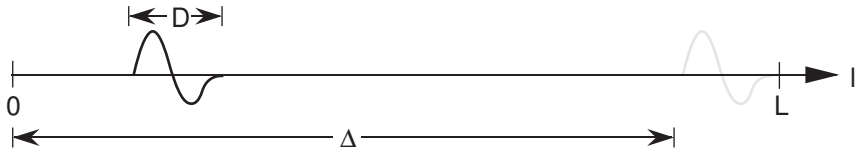


Figure 5.14: Despite uncertainties in the signal's delay Δ, the signal is assumed to lie entirely within the observation interval. Hence the signal's duration D, the duration L of the observation interval, and the maximum expected delay are assumed to be related by $\max\Delta + D - 1 < L$. The figure shows a signal falling properly within the allowed window and a grey one falling just outside.

hood ratio test for detection problems is expressed by

$$\boxed{\Lambda(\mathbf{X}) = \frac{\max_{\boldsymbol{\theta}} p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_1(\boldsymbol{\theta}))}{\max_{\boldsymbol{\theta}} p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_0(\boldsymbol{\theta}))} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \eta}\,.$$

Again, the use of *separate* estimates for each model (rather than for the likelihood ratio as a whole) must be stressed. Unknown signal-related parameter problems and unknown noise-parameter problems have different characteristics; the signal may not be present in one of the observation models. This simplification allows a threshold to be established objectively. In contrast, the noise is present in each model; establishing a threshold value objectively will force new techniques to be developed. We first continue our adventure in unknown-signal-parameter problems, deferring the more challenging unknown-noise-parameter ones to §5.5.2 {192}.

Unknown Signal Delay

A uniformly most powerful decision rule may not exist when an unknown parameter appears in a nonlinear way in the signal model. Most pertinent to array processing is the unknown time origin case: the signal has been subjected to an unknown delay ($s(l-\Delta)$, $\Delta = ?$) and we must determine the signal's presence. The likelihood ratio cannot be manipulated so that the sufficient statistic can be computed without having a value for Δ. Thus, the search for a uniformly most powerful test ends in failure and other methods must be sought. As expected, we resort to the generalized likelihood ratio test.

More specifically, consider the binary test where a signal is either present ($\mathscr{M}_1$) or not ($\mathscr{M}_0$). The signal waveform is known, but its time origin is not. For all possible values of Δ, the delayed signal is assumed to lie *entirely* in the observations (see Fig. 5.14). This signal model is ubiquitous in active sonar and radar, where the reflected signal's exact time-of-arrival is not known and we want to determine whether a return is present or not *and* the value of the delay.* Additive white Gaussian noise is assumed present. The conditional density of the observations under $\mathscr{M}_1$ is

$$p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1;\Delta) = \frac{1}{(2\pi\sigma^2)^{L/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{l=0}^{L-1}[X(l) - s(l-\Delta)]^2\right\}.$$

The exponent contains the only portion of this conditional density that depends on the unknown quantity Δ. Maximizing the conditional density with respect to Δ is equivalent to maximizing

$$\sum_{l=0}^{L-1}\left[X(l)s(l-\Delta) - \frac{1}{2}s^2(l-\Delta)\right].$$

As the signal is assumed to be contained entirely in the observations for all possible values of Δ, the second term does not depend on Δ and equals half of the signal energy E. Rather than analytically maximizing the first term now, we simply write the logarithm of the generalized likelihood ratio test as

$$\max_{\Delta} \sum_{l=\Delta}^{\Delta+D-1} X(l)s(l-\Delta) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta + \frac{E}{2},$$

*For a much more realistic (and harder) version of the active radar/sonar problem, see Problem 5.58.

where the non-zero portion of the summation is expressed explicitly. Using the matched filter interpretation of the sufficient statistic, this decision rule is expressed by

$$\max_{\Delta}[X(l) \otimes s(D-1-l)]\Big|_{l=D-1+\Delta} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma$$

This formulation suggests that the matched filter having a unit-sample response equal to the zero-origin signal be evaluated for each possible value of Δ and that we use the maximum value of the resulting output in the decision rule. In the known-delay case, the matched-filter output is sampled at the "end" of the signal; here, the filter, which has a duration D less than the observation interval L, is allowed to continue processing over the allowed values of signal delay with the maximum output value chosen. The result of this procedure is illustrated in Fig. 5.11 {179}. There two signals, each having the same energy, are passed through the appropriate matched filter. Note that the index at which the maximum output occurs is the maximum likelihood estimate of Δ. Thus, *the detection and the estimation problems are solved simultaneously*. Furthermore, *the amplitude of the signal need not be known* as it enters in expression for the sufficient statistic in a linear fashion and an UMP test exists in that case. We can easily find the threshold γ by establishing a criterion on the false-alarm probability; the resulting simple computation of γ can be traced to the lack of a signal-related quantity or an unknown parameter appearing in $\mathscr{M}_0$.

We have argued the doubtfulness of assuming that the noise is white in discrete-time detection problems. The approach for solving the colored noise problem is to use spectral detection. Handling the unknown delay problem in this way is relatively straightforward. Since a sequence can be represented equivalently by its values or by its DFT, maximization can be calculated in either the time or the frequency domain without affecting the final answer. Thus, the spectral detector's decision rule for the unknown delay problem is

$$\boxed{\max_{\Delta} \sum_{k=0}^{L-1} \left(\frac{\mathrm{Re}[X^*(k)S(k)e^{-j2\pi k\Delta/L}]}{\sigma_k^2} - \frac{1}{2}\frac{|S(k)|^2}{\sigma_k^2} \right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma},$$

where, as usual in unknown delay problems, the observation interval captures the entire signal waveform no matter what the delay might be. The energy term is a constant and can be incorporated into the threshold. The maximization amounts to finding the best linear phase fit to the observations' spectrum once the signal's phase has been removed. A more interesting interpretation arises by noting that the sufficient statistic is itself a Fourier Transform; the maximization amounts to finding the location of the maximum of a sequence given by

$$\mathrm{Re}\left[\sum_{k=0}^{L-1} \frac{X^*(k)S(k)}{\sigma_k^2} e^{-j2\pi k\Delta/L} \right].$$

The spectral detector thus becomes a succession of two Fourier Transforms with the final result determined by the maximum of a sequence!

Unfortunately, the solution to the unknown-signal-delay problem in either the time or frequency domains is confounded when two or more signals are present. Assume two signals are known to be present in the array output, each of which has an unknown delay: $X(l) = s_1(l-\Delta_1) + s_2(l-\Delta_2) + N(l)$. Using arguments similar those used in the one-signal case, the generalized likelihood ratio test becomes

$$\max_{\Delta_1,\Delta_2} \sum_{l=0}^{L-1} X(l)s_1(l-\Delta_1) + X(l)s_2(l-\Delta_2) - s_1(l-\Delta_1)s_2(l-\Delta_2) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta + \frac{E_1+E_2}{2}.$$

Not only do matched filter terms for each signal appear, but also a cross-term between the two signals. It is this latter term that complicates the multiple signal problem: if this term is not zero for *all* possible delays, a non-separable maximization process results and both delays must be varied in concert to locate the maximum. If, however, the two signals are orthogonal regardless of the delay values, the delays can be found separately and the structure of the single signal detector (modified to include matched filters for each signal) will suffice.

This seemingly impossible situation can occur, at least approximately. Using Parseval's Theorem, the cross term can be expressed in the frequency domain.

$$\sum_{l=0}^{L-1} s_1(l-\Delta_1)s_2(l-\Delta_2) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_1(\omega)S_2^*(\omega)e^{j\omega\cdot(\Delta_2-\Delta_1)}\,d\omega$$

For this integral to be zero for all Δ_1, Δ_2, the product of the spectra must be zero. Consequently, if the two signals have disjoint spectral support, they are orthogonal no matter what the delays may be.* Under these conditions, the detector becomes

$$\max_{\Delta_1}\left[X(l)\otimes s_1(D-1-l)\right]\Big|_{l=D-1+\Delta_1} + \max_{\Delta_2}\left[X(l)\otimes s_2(D-1-l)\right]\Big|_{l=D-1+\Delta_2} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma$$

with the threshold again computed independently of the received signal amplitudes.†

$$P_F = Q\left(\frac{\gamma}{\sqrt{(E_1+E_2)\sigma^2}}\right)$$

This detector has the structure of two parallel, independently operating, matched filters, each of which is tuned to the specific signal of interest.

Reality is insensitive to mathematically simple results. The orthogonality condition on the signals that yielded the relatively simple two-signal, unknown-delay detector is often elusive. The signals often share similar spectral supports, thereby violating the orthogonality condition. In fact, we may be interested in detecting the *same* signal repeated twice (or more) within the observation interval. Because of the complexity of incorporating inter-signal correlations, which are dependent on the relative delay, the idealistic detector is often used in practice. In the repeated signal case, the matched filter is operated over the entire observation interval and the number of *excursions* above the threshold noted. An excursion is defined to be a portion of the matched filter's output that exceeds the detection threshold over a contiguous interval. Because of the signal's non-zero duration, the matched filter's response to just the signal has a non-zero duration, implying that the threshold can be crossed at more than a single sample. When one signal is assumed, the maximization step automatically selects the peak value of an excursion. As shown in lower panels of Fig. 5.11 {179}, a low-amplitude excursion may have a peak value less than a non-maximal value in a larger excursion. Thus, when considering multiple signals, the important quantities are the times at which excursion peaks occur, not all of the times the output exceeds the threshold.

Fig. 5.11 illustrates the two kinds of errors prevalent in multiple signal detectors. In the left panel, we find two excursions, the first of which is due to the signal, the second due to noise. This kind of error cannot be avoided; we never said that detectors could be perfect! The right panel illustrates a more serious problem: the threshold is crossed by four excursions, all of which are due to a single signal. Hence, excursions must be sorted through, taking into account the nature of the signal being sought. In the example, excursions surrounding a large one should be discarded if they occur in close proximity. This requirement means that closely spaced signals cannot be distinguished from a single one.

Unknown Signal Waveform

The most general unknown signal parameter problem occurs when the signal itself unknown. This phrasing of the detection problem can be applied to two different sorts of situations. The signal's waveform may not be known precisely because of propagation effects (rapid multipath, for example) or because of source uncertainty. Another situation is the "Hello, is anyone out there?" problem: you want to determine if any non-noise-like quantity is present in the observations. These problems impose severe demands on the detector,

*We stated earlier that this situation happens "at least approximately." Why the qualification?

†Not to be boring, but we emphasize that E_1 and E_2 are the energies of the signals $s_1(l)$ and $s_2(l)$ used in the detector, *not* those of their received correlates $A_1 s_1(l)$ and $A_2 s_2(l)$.

which must function with little *a priori* knowledge of the signal's structure. Consequently, we cannot expect superlative performance.

$$\mathscr{M}_0 : X(l) = N(l)$$
$$\mathscr{M}_1 : X(l) = s(l) + N(l), \quad s(l) = ?$$

The noise is assumed to be Gaussian with a covariance matrix $\mathbf{K}$. The conditional density under $\mathscr{M}_1$ is given by

$$p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1;\mathbf{s}) = \frac{1}{\sqrt{\det[2\pi\mathbf{K}]}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\mathbf{s})^t\mathbf{K}^{-1}(\mathbf{X}-\mathbf{s})\right\}.$$

Using the generalized likelihood ratio test, the maximum value of this density with respect to the unknown "parameters"—the signal values—occurs when $\mathbf{X} = \mathbf{s}$.

$$\max_{\mathbf{s}} p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1;\mathbf{s}) = \frac{1}{\sqrt{\det[2\pi\mathbf{K}]}}$$

The other model does not depend on the signal and the generalized likelihood ratio test for the unknown signal problem, often termed the *square-law* detector, is

$$\boxed{\mathbf{X}^t\mathbf{K}^{-1}\mathbf{X} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma}.$$

For example, if the noise were white, the sufficient statistic is the sum of the squares of the observations.

$$\sum_{l=0}^{L-1} X^2(l) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma$$

If the noise is not white, the detection problem can be formulated in the frequency domain, where the decision rule becomes

$$\sum_{k=0}^{L-1} \frac{|X(k)|^2}{\sigma_k^2} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma.$$

Computation of the false-alarm probability in, for example, the white noise case is relatively straightforward. The probability density of the sum of the squares of L statistically independent zero-mean, unit-variance, Gaussian random variables is termed chi-squared with L degrees of freedom {231}: $\chi^2(L)$.

Example

Assume that the additive noise is white and Gaussian, having a variance σ^2. The sufficient statistic $\Upsilon = \sum X^2(l)$ of the square-law detector has the probability density

$$\Upsilon/\sigma^2 \sim \chi^2(L)$$

when no signal is present. The threshold γ for this statistic is established by $\Pr(\chi^2(L) > \gamma/\sigma^2) = P_F$. The probability of detection is found from the density of a non-central chi-squared random variable having $\chi'^2(L,\lambda)$ having L degrees of freedom and "centrality" parameter $\lambda = \sum_{l=0}^{L-1} \mathsf{E}[X^2(l)]$. In this example, $\lambda = E$, the energy of the observed signal. In Fig. 5.13 {187}, the false-alarm probability was set to 10^{-2} and the resulting probability of detection shown as a function of signal-to-noise ratio. Clearly, the inability to specify the signal waveform leads to a significant reduction in performance. In this example, roughly 10 dB more signal-to-noise ratio is required by the square-law detector than the matched-filter detector, which assumes knowledge of the waveform but not of the amplitude, to yield the same performance.

5.5.2 Unknown Noise Parameters

When aspects of the noise, such as the variance or power spectrum, are in doubt, the detection problem becomes more difficult to solve. Although a decision rule can be derived for such problems using the techniques we have been discussing, establishing a rational threshold value is impossible in many cases. The reason for this inability is simple: *all* models depend on the noise, thereby disallowing a computation of a threshold based on a performance probability. The solution is innovative: derive decision rules and accompanying thresholds that do not depend on false-alarm probabilities!

Consider the case where the variance of the noise is not known and the noise covariance matrix is written as $\sigma^2\widetilde{\mathbf{K}}$ where the trace of $\widetilde{\mathbf{K}}$ is normalized to L. The conditional density of the observations under a signal-related model is

$$p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i;\sigma^2) = \frac{1}{(\sigma^2)^{L/2}\sqrt{\det[2\pi\widetilde{\mathbf{K}}]}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{X}-\mathbf{s}_i)^t\widetilde{\mathbf{K}}^{-1}(\mathbf{X}-\mathbf{s}_i)\right\}.$$

Using the generalized-likelihood-ratio approach, the maximum value of this density with respect to σ^2 occurs when

$$\widehat{\sigma}^2_{\mathrm{ML}} = \frac{(\mathbf{X}-\mathbf{s}_i)^t\widetilde{\mathbf{K}}^{-1}(\mathbf{X}-\mathbf{s}_i)}{L}.$$

This seemingly complicated answer is easily interpreted. The presence of $\widetilde{\mathbf{K}}^{-1}$ in the dot product can be considered a whitening filter. Under the i^{th} model, the expected value of the observation vector is the signal. This computation amounts to subtracting the expected value from the observations, whitening the result, then averaging the squared values—the usual form for the estimate of a variance. Using this estimate for each model, the logarithm of the generalized likelihood ratio becomes

$$\frac{L}{2}\ln\left(\frac{(\mathbf{X}-\mathbf{s}_0)^t\widetilde{\mathbf{K}}^{-1}(\mathbf{X}-\mathbf{s}_0)}{(\mathbf{X}-\mathbf{s}_1)^t\widetilde{\mathbf{K}}^{-1}(\mathbf{X}-\mathbf{s}_1)}\right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta.$$

Computation of the threshold remains. Both models depend on the unknown variance. However, a false-alarm probability, for example, can be computed *if* the probability density of the sufficient statistic does *not* depend on the variance of the noise. In this case, we would have what is known as *constant false-alarm rate* or *CFAR* detector [20],[46: p. 317ff]. If a detector has this property, the value of the statistic will not change if the observations are scaled about their presumed mean. Unfortunately, the statistic just derived does not have this property. Let there be no signal under $\mathscr{M}_0$. The scaling property can be checked in this zero-mean case by replacing $\mathbf{X}$ by $c\mathbf{X}$. With this substitution, the statistic becomes $c^2\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{X}/(c\mathbf{X}-\mathbf{s}_1)^t\widetilde{\mathbf{K}}^{-1}(c\mathbf{X}-\mathbf{s}_1)$. The constant c cannot be eliminated and the detector does not have the *CFAR* property. If, however, the amplitude of the signal is also assumed to be in doubt, a *CFAR* detector emerges. Express the signal component of model i as $A\mathbf{s}_i$, where A is an unknown constant. The maximum likelihood estimate of this amplitude under model i is

$$\widehat{A}_{\mathrm{ML}} = \frac{\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}_i}{\mathbf{s}_i^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}_i}.$$

Using this estimate in the likelihood ratio, we find the decision rule for the *CFAR* detector.*

$$\boxed{\frac{L}{2}\ln\left(\frac{\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{X} - \dfrac{\left(\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}_0\right)^2}{\mathbf{s}_0^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}_0}}{\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{X} - \dfrac{\left(\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}_1\right)^2}{\mathbf{s}_1^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}_1}}\right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta}$$

Now we find that when $\mathbf{X}$ is replaced by $c\mathbf{X}$, the statistic is unchanged. Thus, the probability distribution of this statistic does *not* depend on the unknown variance σ^2. In many applications, no signal is assumed present

*Note that $\widetilde{\mathbf{K}}$ is the *normalized* noise covariance matrix.

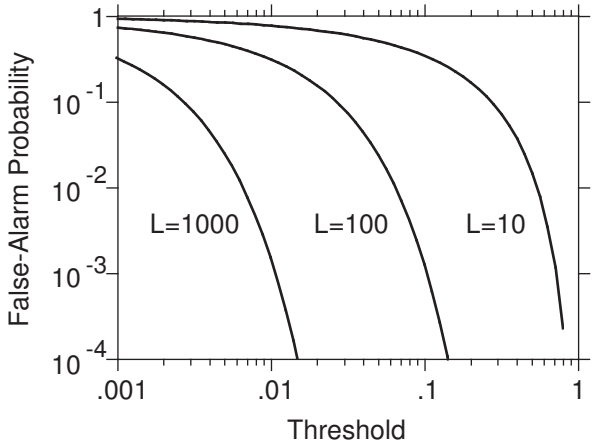


Figure 5.15: The false-alarm probability for the CFAR receiver is plotted against the threshold value γ for several values of L, the number of observations. Note that the test statistic, and thereby the threshold, does not exceed one.

in model $\mathscr{M}_0$; in this case, $\mathscr{M}_0$ does not depend on the unknown amplitude A and a threshold can be found to ensure a specified false-alarm rate for *any* value of the unknown variance. For this specific problem, the likelihood ratio can be manipulated to yield the *CFAR* decision rule

$$\frac{\left(\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}_1\right)^2}{\mathbf{X}^t\widetilde{\mathbf{K}}^{-1}\mathbf{X}\cdot\mathbf{s}_1^t\widetilde{\mathbf{K}}^{-1}\mathbf{s}_1}\underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}}\gamma.$$

Example

Let's extend the previous example to the *CFAR* statistic just discussed to the white noise case. The sufficient statistic is

$$\Upsilon(\mathbf{X})=\frac{[\sum X(l)s(l)]^2}{\sum X^2(l)\cdot\sum s^2(l)}.$$

We first need to find the false-alarm probability as a function of the threshold γ. Using the techniques described in [97], the probability density of $\Upsilon(\mathbf{X})$ under $\mathscr{M}_0$ is given by a Beta density {231}, the parameters of which do *not* depend on either the noise variance (expectedly) or the signal values (unexpectedly).

$$p_{\Upsilon|\mathscr{M}_0}(\Upsilon|\mathscr{M}_0)=\beta\left(\Upsilon,\frac{1}{2},\frac{L-1}{2}\right),$$

We express the false-alarm probability derived from this density as an incomplete Beta function [1], resulting in the curves shown in Fig. 5.15. The statistic's density under model $\mathscr{M}_1$ is related to the non-central F distribution, expressible by the fairly simple, quickly converging, infinite sum of Beta densities

$$p_{\Upsilon|\mathscr{M}_1}(\Upsilon|\mathscr{M}_1)=\sum_{k=0}^{\infty}e^{-d^2}\frac{(d^2)^k}{k!}\beta\left(\Upsilon,k+\frac{1}{2},\frac{L-1}{2}\right),$$

where d^2 equals a signal-to-noise ratio: $d^2=\sum_l s^2(l)/2\sigma^2$. The results of using this *CFAR* detector are shown in Fig. 5.16.

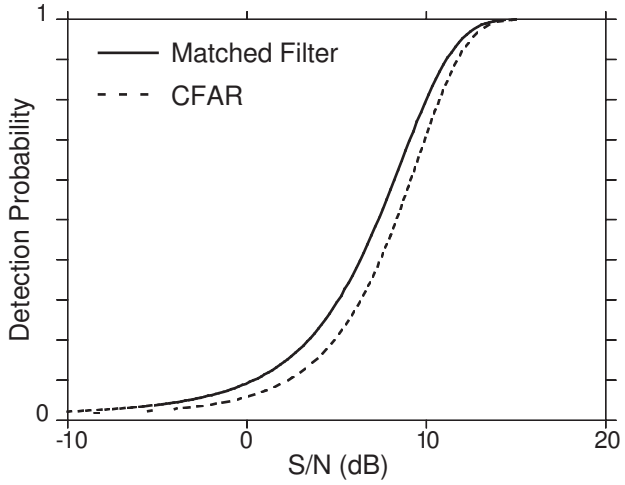


Figure 5.16: The probability of detection for the *CFAR* detector and the matched filter detector is shown as a function of signal-to-noise ratio. The signal and the false-alarm criterion are the same as in Fig. 5.13 {187}. Note how little performance has been lost in this case!

5.6 Non-Gaussian Detection Theory

5.6.1 Partial Knowledge of Probability Distributions

In previous chapters, we assumed we knew the mathematical form of the probability distribution for the observations under each model; some of these distribution's parameters were not known and we developed decision rules to deal with this uncertainty. A more difficult problem occurs when the mathematical form is not known precisely. For example, the data may be approximately Gaussian, containing slight departures from the ideal. More radically, so little may be known about an *accurate* model for the data that we are only willing to assume that they are distributed symmetrically about some value. We develop model evaluation algorithms in this section that tackle both kinds of problems. However, be forewarned that solutions to such general models come at a price: the more specific a model can be that accurately describes a given problem, the better the performance. In other words, the more specific the model, the more the signal processing algorithms can be tailored to fit it with the obvious result that we enhance the performance. However, if our specific model is in error, our neatly tailored algorithms can lead us drastically astray. Thus, the best approach is to relax those aspects of the model which seem doubtful and to develop algorithms that will cope well with worst-case situations should they arise ("And they usually do," echoes every person experienced in the vagaries of data). These considerations lead us to consider nonparametric variations in the probability densities *compatible* with our assessment of model accuracy and to derive decision rules that *minimize* the impact of the worse-case situation.

Worst-Case Probability Distributions

In model evaluation problems, there are "optimally" hard problems, those where the models are the most difficult to distinguish. The impossible problem is to distinguish models that are identical. In this situation, the conditional densities of the observed data are equal and the likelihood ratio is constant for all possible values of the observations. It is obvious that identical models are indistinguishable; this elaboration suggests that in terms of the likelihood ratio, *hard problems are those in which the likelihood ratio is constant*. Thus, "hard problems" are those in which the class of conditional probability densities has a constant ratio for wide ranges of observed data values.

The most relevant model evaluation problem for us is the discrimination between two models that differ only in the means of statistically independent observations: the conditional densities of each observation are related as $p_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1) = p_{X_l|\mathscr{M}_0}(X_l - m|\mathscr{M}_0)$. Densities that would make this model evaluation problem

hard would satisfy the functional equation

$$p(x-m) = C(m) \cdot p(x), \quad \forall\, x \geq m$$

where $C(m)$ is quantity depending on the mean m, but not the variable x.* For probability densities satisfying this equation, any value of the observed datum which has a value greater than m cannot be used to distinguish the two models. If one considers only those zero-mean densities $p(\cdot)$ which are symmetric about the origin, then by symmetry the likelihood ratio would also be constant for $x \leq 0$. Hypotheses having these densities could only be distinguished when the observations lay in the interval $(0,m)$; such model evaluation problems are hard!

From the functional equation, we see that the quantity $C(m)$ must be inversely proportional to $p(m)$ (substitute $x = m$ into the equation). Incorporating this fact into our functional equation, we find that the *only* solution is the exponential function.

$$p(x-m) = C(m) \cdot p(x) \implies p(x) \propto e^{-x}, x \geq 0$$

If we insist that the density satisfying the functional equation be symmetric, the solution is the so-called Laplacian (or double-exponential) density.

$$p_x(x) = \frac{1}{\sqrt{2\sigma^2}} \exp\left\{ -\frac{|x|}{\sqrt{\sigma^2/2}} \right\}$$

When this density serves as the underlying density for our hard model-testing problem, the likelihood ratio has the form [49],[50],[80: pp. 175–187]

$$\ln \Lambda(X_l) = \begin{cases} -\dfrac{m}{\sqrt{\sigma^2/2}}, & X_l < 0 \\ \dfrac{2X_l - m}{\sqrt{\sigma^2/2}}, & 0 < X_l < m \\ \dfrac{m}{\sqrt{\sigma^2/2}}, & m < X_l \end{cases}$$

Indeed, the likelihood ratio is constant over much of the range of values of X_l, implying that the two models are very similar over those ranges. This worst-case result will appear repeatedly as we embark on searching for the model evaluation rules that minimize the effect of modeling errors on performance.

5.6.2 Robust Hypothesis Testing

"Robust" is a technical word that implies insensitivity to modeling assumptions. As we have seen, some algorithms are robust while others are not. The intent of *robust signal processing* to derive algorithms that are *explicitly* insensitive to the underlying signal and/or noise models. The way in which modeling uncertainties are described is typified by the approach we shall use in the following discussion of robust model evaluation.

We assume that two *nominal* models of the generation of the statistically independent observations are known; the "actual" conditional probability density that describes the data under the assumptions of each model is not known exactly, but is "close" to the nominal. Letting $p(\cdot)$ be the actual probability density for each observation and $p^o(\cdot)$ the nominal, we say that [50]

$$p(x) = (1-\varepsilon)p^o(x) + \varepsilon p^d(x)\,,$$

where p^d is the unknown disturbance density and ε is the uncertainty variable $(0 \leq \varepsilon < 1)$. The uncertainty variable specifies how accurate the nominal model is thought to be: the smaller ε, the smaller the contribution of the disturbance. It is assumed that some value for ε can be rationally assigned. The disturbance density is entirely unknown and is assumed to be *any* valid probability density function. The expression given above is normalized so that $p(\cdot)$ has unit area. Thus, in this description of modeling uncertainties, the nominal density sits at the "center", with the actual density ranging about it. An example of densities described this way are shown in Fig. 5.17.

*The uniform density does not satisfy this equation as the domain of the function $p(\cdot)$ is assumed to be infinite.

The robust model evaluation problem is formally stated as

$$\mathscr{M}_0 \colon p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{x}|\mathscr{M}_0) = \prod_{l=0}^{L-1} \left[(1-\varepsilon) p^o_{X_l|\mathscr{M}_0}(X_l|\mathscr{M}_0) + \varepsilon p^d_{X_l|\mathscr{M}_0}(X_l|\mathscr{M}_0) \right]$$

$$\mathscr{M}_1 \colon p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{x}|\mathscr{M}_1) = \prod_{l=0}^{L-1} \left[(1-\varepsilon) p^o_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1) + \varepsilon p^d_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1) \right]$$

The nominal densities under each model correspond to the conditional densities that we have been using until now. The disturbance densities are intended to model imprecision of both descriptions; hence, they are assumed to be different in the context of each model. Note that the measure of imprecision ε is assumed to be the same under either model.

To solve this problem, we take what is known as a *minimax* approach: find the worst-case combinations of *a priori* densities (max), then minimize the consequences of this situation (mini) according to some criterion. In this way, bad situations are handled as well as can be expected while the more tolerable ones are (hopefully) processed well also. The "mini" phase of the minimax solution corresponds to the likelihood ratio for many criteria. Thus, the "max" phase amounts to finding the worst-case probability distributions for the likelihood ratio test as described in the previous section: find the disturbance densities that can result in a constant value for the ratio over large domains of the likelihood ratio. This evaluation is most easily understood graphically by considering the probability *distribution* functions. When the two nominal distributions scaled by $1-\varepsilon$ can be brought together so that they are equal for some disturbance, then the likelihood ratio will be constant in that domain. Of most interest here is the case where the models differ only in the value of the mean, as shown in Fig. 5.18. "Bringing the distributions together" means, in this case, scaling the distribution for $\mathscr{M}_0$ by $1-\varepsilon$ while adding the constant ε to the scaled distribution for $\mathscr{M}_1$. One can shown in general that if the ratio of the nominal densities is monotonic, this procedure finds the worst-case distribution [49]. The distributions overlap for small and for large values of the data with no overlap in a central region. As we shall see, the size of this central region depends greatly on the choice of ε. The *tails* of the worst-case distributions under each model are equal; conceptually, we consider that the worst-case densities have exponential tails in the model evaluation problem.

Letting p^w denote the worst-case density, our minimax procedure results in the following densities for each model in the likelihood ratio test.

$$p^w_{X_l|\mathscr{M}_i}(X_l|\mathscr{M}_i) = \begin{cases} p^o_{X'_l|\mathscr{M}_0}(X'_l|\mathscr{M}_0) C'_i e^{-K'|X_l - X'_l|} & X_l < X'_l \\ p^o_{X_l|\mathscr{M}_i}(X_l|\mathscr{M}_i) & X'_l < X_l < X''_l \\ p^o_{X''_l|\mathscr{M}_0}(X''_l|\mathscr{M}_0) C''_i e^{-K''|X_l - X''_l|} & X_l > X''_l \end{cases}$$

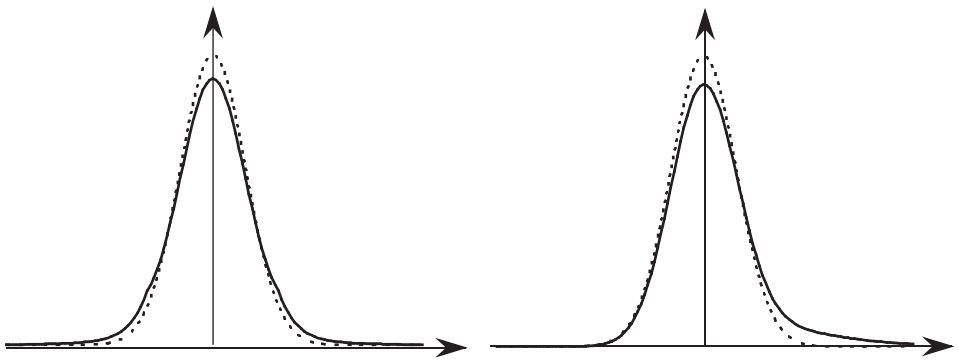


Figure 5.17: The nominal density, a Gaussian, is shown as a dashed line along with example densities derived from it having an uncertainty of 10% ($\varepsilon = 0.1$). The left plot illustrates a symmetric contamination and the right an asymmetric one.

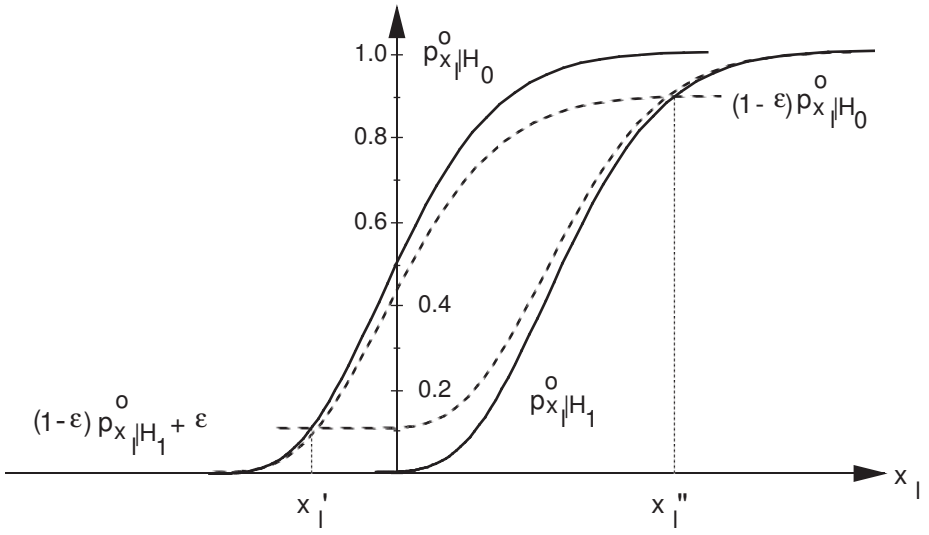


Figure 5.18: Nominal probability distributions for each model are shown. The worst-case distributions corresponding to these are also shown for the uncertainty variable ε equaling 0.1.

The constants K' and K'' determine the rate of decay of the exponential tails of these worst-case distributions. Their specific values have not yet been determined, but since they are not needed to compute the likelihood ratio, we don't need them. The constants C_i' and C_i'' are required so that a unit-area density results. The likelihood ratio for each observation in the robust model evaluation problem becomes

$$\Lambda(X_l) = \begin{cases} \dfrac{C_1'}{C_0'} & X_l < X_l' \\ \dfrac{p^o_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1)}{p^o_{X_l|\mathscr{M}_0}(X_l|\mathscr{M}_0)} & X_l' < X_l < X_l'' \\ \dfrac{C_1''}{C_0''} & X_l'' < X_l\,. \end{cases}$$

The evaluation of the likelihood ratio depends entirely on determining values for X_l' and X_l''. The ratios $C_1'/C_0' = c'$ and $C_1''/C_0'' = c''$ are easily found; in the tails, the value of the likelihood ratio equals that at the edges of the central region for continuous densities.

$$c' = \frac{p^o_{X_l|\mathscr{M}_1}(X_l'|\mathscr{M}_1)}{p^o_{X_l|\mathscr{M}_0}(X_l'|\mathscr{M}_0)} \quad c'' = \frac{p^o_{X_l|\mathscr{M}_1}(X_l''|\mathscr{M}_1)}{p^o_{X_l|\mathscr{M}_0}(X_l''|\mathscr{M}_0)}$$

At the left boundary, for example, the distribution functions must satisfy $(1-\varepsilon)P_{X_l|\mathscr{M}_0}(X_l'|\mathscr{M}_0) = (1-\varepsilon)P_{X_l|\mathscr{M}_1}(X_l'|\mathscr{M}_1)+\varepsilon$. In terms of the nominal densities, we have

$$\int_{-\infty}^{X_l'} \left[p_{X_l|\mathscr{M}_0}(x|\mathscr{M}_0) - p_{X_l|\mathscr{M}_1}(x|\mathscr{M}_1)\right]\,dx = \frac{\varepsilon}{1-\varepsilon}.$$

This equation also applies to the value right edge X_l''. Thus, for a given value of ε, the integral of the difference between the nominal densities should equal the ratio $\varepsilon/(1-\varepsilon)$ for two values. Fig. 5.19 illustrates this effect for a Gaussian example. The bi-valued nature of this integral may not be valid for some values of ε; the value chosen for ε can be too large, making it impossible distinguish the models! This unfortunate circumstance means that the uncertainties, as described by the value of ε, swamp the characteristics that distinguish the

models. Thus, the models must be made more precise (more must be known about the data) so that smaller deviations from the nominal models can describe the observations.

Returning to the likelihood ratio, the "robust" decision rule consists of computing a *clipped* function of each observed value, multiplying them together, and comparing the product computed over the observations with a threshold value. We assume that the nominal distributions of each of the L observations are equal; the values of the boundaries X_l' and X_l'' then do not depend on the observation index l in this case. More simply, evaluating the logarithm of the quantities involved results in the decision rule

$$\sum_{l=0}^{L-1} f(X_l) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma.$$

where the function $f(\cdot)$ is the clipping function given by

$$f(X_l) = \begin{cases} \ln c' & X_l < X' \\ \ln \dfrac{p^o_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1)}{p^o_{X_l|\mathscr{M}_0}(X_l|\mathscr{M}_0)} & X' < X_l < X'' \\ \ln c'' & X'' < X_l \,. \end{cases}$$

If the observations were not identically distributed, then the clipping function would depend on the observation index.*

Determining the threshold γ that meets a specific performance criterion is difficult in the context of robust model evaluation. By the very nature of the problem formulation, some degree of uncertainty in the *a priori* densities exists. A specific false-alarm probability can be guaranteed by using the worst-case distribution under $\mathscr{M}_0$. This density has the disturbance term being an impulse at infinity. Thus, the expected value m_c of a clipped observation $f(X_l)$ with respect to the worst-case density is $(1-\varepsilon)\,\mathsf{E}[f(X_l)] + \varepsilon \ln c''$ where the expected value in this expression is evaluated with respect to the nominal density under $\mathscr{M}_0$. Similarly, an expression for the variance σ_c^2 of the clipped observation can be derived. As the decision rule computes the sum of the clipped, statistically independent observations, the Central Limit Theorem can be applied to the sum, with the result that the worst-case false-alarm probability will approximately equal $Q\big((\gamma - Lm_c)/(\sqrt{L}\sigma_c)\big)$. The threshold γ can then be found which will guarantee a specified performance level. Usually, the worst-case situation does not occur and the threshold set by this method is conservative. We can assess the degree of conservatism by evaluating these quantities under the nominal density rather than the worst-case density.

Example

Let's consider the Gaussian model evaluation problem we have been using so extensively. The individual observations are statistically independent and identically distributed with variance five: $\sigma^2 = 5$. For model $\mathscr{M}_0$, the mean is zero; for $\mathscr{M}_1$, the mean is one. These nominal densities describe our best models for the observations, but we seek to allow slight deviations (10%) from them. The equation to be solved for the boundaries is the implicit equation

$$Q\left(\frac{z-m}{\sigma}\right) - Q\left(\frac{z}{\sigma}\right) = \frac{\varepsilon}{1-\varepsilon}.$$

The quantity on the left side of the equation is shown in Fig. 5.19. If the uncertainty in the Gaussian model, as expressed by the parameter ε, is larger than 0.15 (for the example values of m and σ), no solution exists. Assuming that ε equals 0.1, the quantity $\varepsilon/(1-\varepsilon) = 0.11$ and the clipping thresholds are $X' = -1.675$ and $X'' = 2.675$. Between these values, the clipping function is given by the logarithm of the likelihood ratio, which is given by $(2mX_l - m^2)/(2\sigma^2)$.

We can decompose the clipping operation into a cascade of two operations: a linear scaling and shifting (as described by the previous expression) followed by a clipper having unit slope (see

*Note that we only need to require that the *nominal* density remain constant throughout the observations. The disturbance density, and through it the density of each observation, could vary without disturbing the validity of this result! Such generality is typical when one loosens modeling restrictions, but, as we have said, this generality is bought with diminished performance.

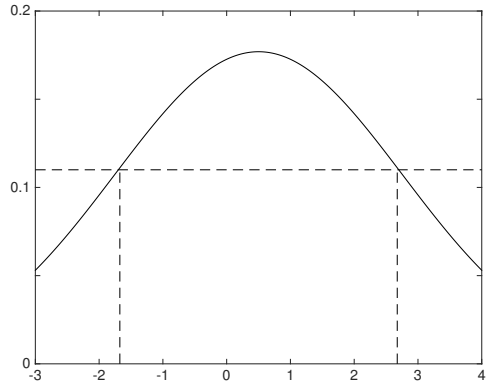


Figure 5.19: The quantity used to determine the thresholds in the robust decision rule is shown when $m = 1$ and $\sigma^2 = 5$. Given a value of ε, a value on the vertical axis is selected and the corresponding values on the horizontal axis yield the thresholds.

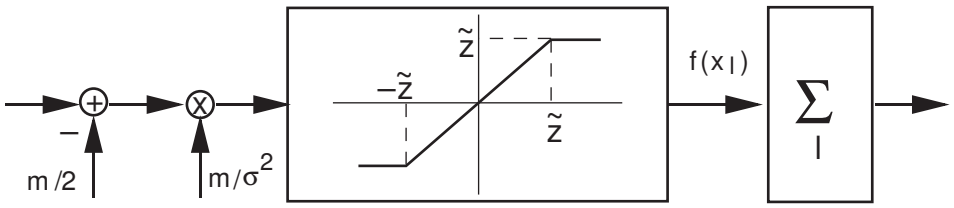


Figure 5.20: The robust decision rule for the case of a Gaussian nominal density is shown. The observations are first scaled and shifted by quantities that depend on the mean m and the variance σ^2. The resulting quantity is then passed through a symmetric unit-slope clipping function whose clipping thresholds also depend on the parameters of the distributions.

Fig. 5.20). Let $\tilde{X}_l$ denote the result of the scaling and shifting operation. This quantity has mean $m^2/2\sigma^2$ and variance m^2/σ^2 under $\mathscr{M}_1$ and the opposite signed mean and the same variance under $\mathscr{M}_0$. The threshold values of the unit-clipping function are thus given by the solution of the equation

$$Q\left(\frac{\tilde{z}+\dfrac{m^2}{2\sigma^2}}{m/\sigma}\right) - Q\left(\frac{\tilde{z}-\dfrac{m^2}{2\sigma^2}}{m/\sigma}\right) = \frac{\varepsilon}{1-\varepsilon}\,.$$

By substituting $-\tilde{z}$ for $\tilde{z}$ in this equation, we find that the two solutions are negatives of each other. We have now placed the unit-clipper's threshold values symmetrically about the origin; however, they do depend on the value of the mean m. In this example, the threshold is numerically given by $\tilde{z} = 0.435$. The expected value of the result of the clipping function with respect to the worst-case density is given by the complicated expression

$$\mathsf{E}[f(X_l)] = (1-\varepsilon)\left[X'Q\left(\frac{-X'}{\sigma}\right) + X''Q\left(\frac{X''}{\sigma}\right) + \sqrt{\frac{\sigma^2}{2\pi}}\left(e^{-\frac{X'^2}{2\sigma^2}} - e^{-\frac{X''^2}{2\sigma^2}}\right)\right] + \varepsilon X''$$

The variance is found in a similar fashion and can be used to find the threshold γ on the sum of clipped observation values.

5.6.3 Non-Parametric Model Evaluation

If *very* uncertain about model accuracy, assuming a form for the nominal density may be questionable or quantifying the degree of uncertainty may be unreasonable. In these cases, *any* formula for the underlying probability densities may be unjustified, but the model evaluation problem remains. For example, we may want to determine a signal's presence or absence in a radar's output (non-zero mean vs. zero mean) without much knowledge of the contaminating noise. If minimal assumptions can be made about the probability densities, *non-parametric model evaluation* can be used [38]. In this theoretical framework, no formula for the conditional densities is required; instead, we use worst-case densities which conform to the weak problem specification. Because few assumptions about the probability models are used, non-parametric decision rules are robust: they are insensitive to modeling assumptions because so few are used. The "robust" test of the previous section are so-named because they explicitly encapsulate model imprecision. In either case, one should expect greater performance (smaller error probabilities) in non-parametric decision rules than possible from a "robust" one.

Two hypothesized models are to be tested; $\mathscr{M}_0$ is intended to describe the situation "the observed data have zero mean" and the other "a non-zero mean is present". We make the usual assumption that the L observed data values are statistically independent. The *only* assumption we will make about the probabilistic descriptions underlying these models is that the median of the observations is zero in the first instance and non-zero in the second. The *median* of a random variable is the "half-way" value: the probability that the random variable is less than the median is one-half as is the probability that it is greater. The median and mean of a random variable are not necessarily equal; for the special case of a symmetric probability density they are. In any case, the non-parametric models will be stated in terms of the probability that an observation is greater than zero.

$$\mathscr{M}_0 : \Pr[X_l \geq 0] = \tfrac{1}{2}$$
$$\mathscr{M}_1 : \Pr[X_l \geq 0] > \tfrac{1}{2}$$

The first model is equivalent to a zero-median model for the data; the second implies that the median is greater than zero. Note that the form of the two underlying probability densities need not be the same to correspond to the two models; they can differ in more general ways than in their means.

To solve this model evaluation problem, we seek (as do all robust techniques) the worst-case density, the density satisfying the conditions for one model that is maximally difficult to distinguish from a given density under the other. Several interesting problems arise in this approach. First of all, we seek a non-parametric answer: the solution must not depend on unstated parameters (we should not have to specify how large the non-zero mean might be). Secondly, the model evaluation rule must not depend on the form for the given density. These seemingly impossible properties are easily satisfied. To find the worst-case density, first define $p^{+}_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1)$ to be the probability density of the l^{th} observation assuming that $\mathscr{M}_1$ is true *and* that the observation was non-negative. A similar definition for negative values is needed.

$$p^{+}_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1) = p_{X_l|\mathscr{M}_1, X_l \geq 0}(X_l|\mathscr{M}_1, X_l \geq 0)$$
$$p^{-}_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1) = p_{X_l|\mathscr{M}_1, X_l < 0}(X_l|\mathscr{M}_1, X_l < 0)$$

In terms of these quantities, the conditional density of an observation under $\mathscr{M}_1$ is given by

$$p_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1) = \Pr[X_l \geq 0|\mathscr{M}_1] p^{+}_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1) + (1 - \Pr[X_l \geq 0|\mathscr{M}_1]) p^{-}_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1)$$

The worst-case density under $\mathscr{M}_0$ would have *exactly* the same functional form as this one for positive and negative values while having a zero median.* As depicted in Fig. 5.21, a density meeting these requirements is

$$p_{X_l|\mathscr{M}_0}(X_l|\mathscr{M}_0) = \frac{p^{+}_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1) + p^{-}_{X_l|\mathscr{M}_1}(X_l|\mathscr{M}_1)}{2} .$$

The likelihood ratio for a single observation would be $2\Pr[X_l \geq 0|\mathscr{M}_1]$ for non-negative values and $2(1 - \Pr[X_l \geq 0|\mathscr{M}_1])$ for negative values. While the likelihood ratio depends on $\Pr[X_l \geq 0|\mathscr{M}_1]$, which is

*Don't forget that the worst-case density in model evaluation agrees with the given one over as large a range as possible.

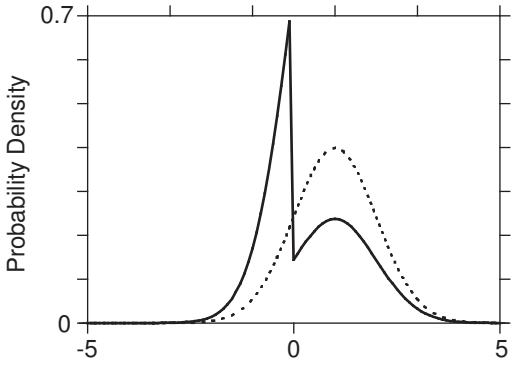


Figure 5.21: For each density having a positive-valued median, the worst-case density having zero median would have exactly the same functional form as the given one on the positive and negative real lines, but with the areas adjusted to be equal. Here, a unit-mean, unit-variance Gaussian density and its corresponding worst-case density are shown. The resulting worst-case density is usually discontinuous at the origin; be that as it may, this rather bizarre worst-case density leads to a simple non-parametric decision rule.

not specified in our non-parametric model, the sufficient statistic will *not* depend on it! To see this, note that likelihood ratio varies only with the sign of the observation. Hence, the optimal decision rule amounts to counting how many of the observations are positive; this count can be succinctly expressed with the unit-step function $\mathrm{u}(\cdot)$ as $\sum_{l=0}^{L-1}\mathrm{u}(X_l)$.[†] Thus, the likelihood ratio for the L statistically independent observations is written

$$2^L\cdot\Pr[X_l\geq 0|\mathscr{M}_1]^{\sum_l \mathrm{u}(X_l)}\cdot(1-\Pr[X_l\geq 0|\mathscr{M}_1])^{L-\sum_l \mathrm{u}(X_l)}\underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}}\eta.$$

Making the usual simplifications, the unknown probability $\Pr[X_l\geq 0|\mathscr{M}_1]$ can be maneuvered to the right side and merged with the threshold. The optimal non-parametric decision rule thus compares the sufficient statistic—the count of positive-valued observations—with a threshold determined by the Neyman-Pearson criterion.

$$\boxed{\sum_{l=0}^{L-1}\mathrm{u}(X_l)\underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}}\gamma}$$

This decision rule is called the *sign test* as it depends only on the signs of the observed data. The sign test is uniformly most powerful and robust.

To find the threshold γ, we can use the Central Limit Theorem to approximate the probability distribution of the sum by a Gaussian. Under $\mathscr{M}_0$, the expected value of $\mathrm{u}(X_l)$ is $\frac{1}{2}$ and the variance is $\frac{1}{4}$. To the degree that the Central Limit Theorem reflects the false-alarm probability (see Problem 5.64), P_F is approximately given by

$$P_F = Q\left(\frac{\gamma-L/2}{\sqrt{L/4}}\right)$$

and the threshold is found to be

$$\gamma=\frac{\sqrt{L}}{2}Q^{-1}(P_F)+\frac{L}{2}.$$

As it makes no sense for the threshold to be greater than L (how many positively values observations can there be?), the specified false-alarm probability must satisfy $P_F\geq Q\left(\sqrt{L}\right)$. This restriction means that increasing stringent requirements on the false-alarm probability can only be met if we have sufficient data.

5.6.4 Partially Known Signals and Noise

Rather than assuming that aspects of the signal, such as its amplitude, are beyond any set of justifiable assumptions and are thus "unknown", we may have a situation where these signal aspects are "uncertain". For example, the amplitude may be known to be within ten percent of a nominal value. If the case, we would

[†]We define the unit-step function as $\mathrm{u}(x)=\begin{cases}1 & x>0\\ 0 & x<0\end{cases}$, with the value at the origin undefined. We presume that the densities have no mass at the origin under either model. Although appearing unusual, $\sum\mathrm{u}(X_l)$ does indeed yield the number of positively valued observations.

expect better performance characteristics from a detection strategy exploiting this partial knowledge from one that doesn't. To derive detectors that use partial information about signal and noise models, we apply the approach used in robust model evaluation: find the worst-case combination of signal and noise consistent with the partial information, then derive the detection strategy that best copes with it. We have seen that the optimal detection strategy is found from the likelihood ratio: no matter what the signal and noise models are, the likelihood ratio yields the best decision rule. When applied to additive Gaussian noise problems, the performance of the likelihood ratio test increases with the signal-to-noise ratio of the difference between the two hypothesized signals. Since we focus on deciding whether a particular signal is present or not, performance is determined by that signal's *SNR* and the worst-case situation occurs when this signal-to-noise ratio is smallest. The results from robust model evaluation taught us to design the detector to the worst-case situation, which in our case roughly means employing matched filters based on the worst-case signal. Employing this approach results in what are known as *robust detectors*.

5.6.5 Partially Known Signal Waveform

The nominal signal waveform is known, but the actual signal present in the observed data can be corrupted slightly. Using the minimax approach {196}, we seek the worst possible signal that could be present consistent with constraints on the corruption. Once we know what signal that is, our detector should consist of a filter matched to that worst-case signal. Let the observed signal be of the form $s(l) = s^o(l) + c(l)$: $s^o(l)$ is the nominal signal and $c(l)$ the corruption in the observed signal. The nominal-signal energy is E^o and the signal corruption is assumed to have an energy that is less than E^c. Spectral techniques are assumed to be applicable so that the covariance matrix of the Fourier transformed noise is diagonal. The signal-to-noise ratio is given by $\sum_k |S(k)|^2/\sigma_k^2$. What corruption yields the smallest signal-to-noise ratio? The worst-case signal will have the largest amount of corruption possible. This constrained minimization problem—minimize the signal-to-noise ratio while forcing the energy of the corruption to be less than E^c—can then be solved using Lagrange multipliers.

$$\min_{\{S_k\}} \left[\sum_k |S(k)|^2/\sigma_k^2 + \lambda \left(\sum_k |C(k)|^2 - E^c\right)\right]$$

By evaluating the appropriate derivatives, we find the spectrum of the worst-case signal to be a frequency-weighted version of the nominal signal's spectrum.

$$S^w(k) = \frac{\lambda\sigma_k^2}{1+\lambda\sigma_k^2} S^o(k) \quad \text{where} \quad \sum_k \left(\frac{1}{1+\lambda\sigma_k^2}\right)^2 |S^o(k)|^2 = E^c$$

The only unspecified parameter is the Lagrange multiplier λ, with the latter equation providing an implicit solution in most cases.

If the noise is white, the σ_k^2 equal a constant, implying that the worst-case signal is a scaled version of the nominal, equaling $S^w(k) = \left(1 - \sqrt{E^c/E^o}\right) S^o(k)$. The robust decision rule derived from the likelihood ratio is given by

$$\text{Re}\left[\sum_{k=0}^{L-1} X^*(k) S^o(k)\right] \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma.$$

By incorporating the scaling constant $\left(1 - \sqrt{E^c/E^o}\right)$ into the threshold γ, we find that *the matched filter used in the white noise, known-signal case is robust with respect to signal uncertainties*. The threshold value is identical to that derived using the nominal signal as model $\mathscr{M}_0$ does not depend on the uncertainties in the signal model. Thus, the detector used in the known-signal, white noise case can also be used when the signal is partially corrupted. Note that in solving the general signal corruption case, the imprecise signal amplitude situation was also solved.

If the noise is not white, the proportion between the nominal and worst-case signal spectral components is not a constant. The decision rule expressed in terms of the frequency-domain sufficient statistic becomes

$$\text{Re}\left[\sum_{k=0}^{L-1} \left(\frac{\lambda\sigma_k^2}{1+\lambda\sigma_k^2}\right) X^*(k) S^o(k)\right] \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma.$$

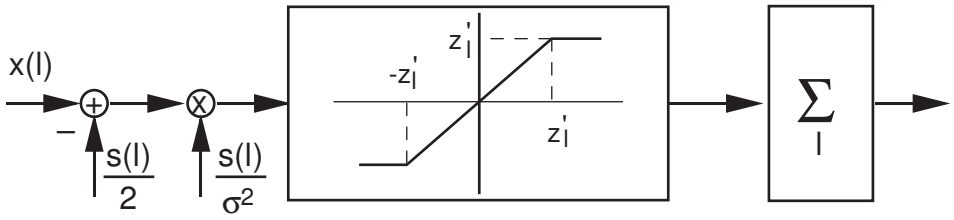


Figure 5.22: The robust detector consists of a linear scaling and shifting operation followed by a unit-slope clipper, whose clipping thresholds depend on the value of the signal. The key element is the clipper, which serves to censor large excursions of the observed from the value of the signal.

Thus, the detector derived for colored noise problems is *not* robust. The threshold depends on the noise spectrum and the energy of the corruption. Furthermore, calculating the value of the Lagrange multiplier in the colored noise problem is quite difficult, with multiple solutions to its constraint equation quite possible. Only one of these solutions will correspond to the worst-case signal.

5.6.6 Partially Known Noise Amplitude Distribution

The previous sections assumed that the probability distribution of the noise was known precisely, and furthermore, that this distribution was Gaussian. Deviations from this assumption occur frequently in applications [68, 73, 74]. We *expect* Gaussian noise in situations where noise sources are many and of roughly equal strength: the Central Limit Theorem suggests that if these noise sources are independent (or even mildly dependent), their superposition will be Gaussian. As shown in §2.1.10 {11}, the Central Limit Theorem converges very slowly and deviations from this model, particularly in the tails of the distribution, are a fact of life. Furthermore, unexpected, deviant noise sources also occur and these distort the noise amplitude distribution. Examples of this phenomenon are lightning (causing momentary, large, skewed changes in the nominal amplitude distribution in electromagnetic sensitive sensors) and ice fractures in polar seas (evoking similar distributional changes in acoustic noise). These changes are momentary and their effects on the amplitude distribution are, by and large, unpredictable. For these situations, we invoke ideas from robust model evaluation to design robust detectors, ones insensitive to deviations from the Gaussian model [33],[58],[80: pp. 175–187].

We assume that the noise component in the observations consists of statistically independent elements, each having a probability amplitude density of the form

$$p_N(N) = (1-\varepsilon)p_N^o(N) + \varepsilon p_N^d(N)$$

where $p_N^o(\cdot)$ is the nominal noise density, taken to be Gaussian, and $p_N^d(\cdot)$ is the deviation of the actual density from the nominal, also a density. This ε-*contamination* model [50] is parameterized by ε, the uncertainty variable, a positive number less than one that defines how large the deviations from the nominal can be. As shown in §5.6.2 {195}, the decision rule for the robust detector is

$$\sum_{l=0}^{L-1} f_l\left[\frac{X(l)s(l) - s^2(l)/2}{\sigma^2}\right] \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma,$$

where $f_l(\cdot)$ is a memoryless nonlinearity having the form of a clipper {198}. The block diagram of this receiver is shown diagrammatically in Fig. 5.22. The clipping function threshold z_l' is related to the assumed variance σ^2 of the nominal Gaussian density, the deviation parameter ε, *and* the signal value at the l^{th} sample by the positive-valued solution of

$$Q\left(\frac{z_l' + \dfrac{s^2(l)}{2\sigma^2}}{s(l)/\sigma}\right) - Q\left(\frac{z_l' - \dfrac{s^2(l)}{2\sigma^2}}{s(l)/\sigma}\right) = \frac{\varepsilon}{1-\varepsilon}.$$

An example of solving for ε is shown in Fig. 5.19 {199}.

The characteristics of the clipper vary with each signal value: the clipper is exquisitely tuned to the proper signal value, ignoring values that deviate from the signal. Furthermore, note that this detector relies on *large* signal values relative to the noise. If the signal values are small, the above equation for z_l' has *no* solution. Essentially, the robust detector ignores those values since the signal is too weak to prevent the noise density deviations from entirely confusing the models.* The threshold can be established for the signal values used by the detector through the Central Limit Theorem as described in §5.6.2 {195}.

5.6.7 Non-Gaussian Observations

The term "non-Gaussian" has a connotation similar to that of "non-linear"; rather than meaning those problems which are not simple (the simple ones being the Gaussian or linear ones), these terms refer instead the general problem—all possible stationary random sequences or all systems. In general, "non-Gaussian detection theory" makes *no* assumption as to the specific form of the noise amplitude distribution [57]. This generality mimics situations where the additive noise is variable, having an unpredictable structure which makes *a priori* models of the noise difficult to justify. This section describes detection algorithms that make few assumptions about the joint probability density function of the noise amplitudes at several samples. For simplicity, the noise sequence is assumed in sequel to be white.

Small-Signal Detection

Before we introduce truly non-Gaussian detection strategies, we need to discuss the structure of the detector when the noise amplitude distribution is known. From the results presented while solving the general model evaluation problem, we find that the likelihood ratio test for statistically independent noise values is

$$\sum_{l=0}^{L-1}\{\ln p_N[X(l)-s_1(l)]-\ln p_N[X(l)-s_0(l)]\} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma.$$

The term in braces is Υ_l, the sufficient statistic at the l^{th} sample. As L is usually large and each term in the summation is statistically independent of the others, the threshold γ is found by approximating the distribution of the sufficient statistic by a Gaussian (the Central Limit Theorem again). Computing the mean and variance of each term, the false-alarm probability is found to be

$$P_F = Q\left(\frac{\ln\eta - \sum \mathsf{E}[\Upsilon_l|\mathscr{M}_0]}{\sqrt{\sum \mathsf{var}[\Upsilon_l|\mathscr{M}_0]}}\right).$$

This general result applies to any assumed noise density, including the Gaussian. The matched filter answer derived earlier is contained in the decision rule given above.

A matched-filter-like answer also results when small signal-to-noise ratio problems are considered for *any* amplitude distribution [57: pp. 5–8],[86],[87]. Assume that $\ln p_N\big(X(l)-s(l)\big)$, considered as a function of the signal value $s(l)$, can be expressed in a Taylor series centered at $X(l)$.

$$\ln p_N\big(X(l)-s(l)\big) = \ln p_N\big(X(l)\big) - s(l)\left.\frac{d\ln p_N(N)}{dN}\right|_{N=X(l)} + \frac{1}{2}s^2(l)\left.\frac{d^2\ln p_N(N)}{dN^2}\right|_{N=X(l)} + \cdots$$

In the *small* signal-to-noise ratio case, second and higher order terms are neglected, leaving the decision rule

$$\sum_{l=0}^{L-1}\left\{-s_1(l)\left.\frac{d\ln p_N(N)}{dN}\right|_{N=X(l)} + s_0(l)\left.\frac{d\ln p_N(N)}{dN}\right|_{N=X(l)}\right\} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma.$$

This rule says that in the small signal case, the sufficient statistic for each signal is the result of match-filtering the result of passing the observations through a front-end memoryless non-linearity $-d\ln p_N(x)/dx$, which depends *only* on noise amplitude characteristics (Fig. 5.23).

*In the next section, we present a structure applicable in small signal-to-noise ratio cases.

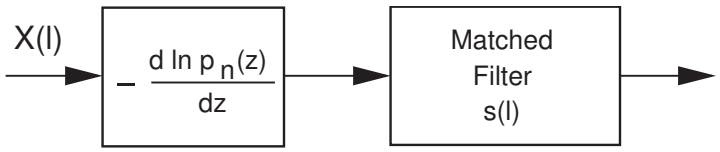


Figure 5.23: The decision rule for the non-Gaussian case when the signal is small can be expressed as a matched filter where the observations are first passed through a memoryless nonlinearity.

Because of the presence of the matched filter and the relationship of the matched filter to Gaussian noise problems, one might presume that the nonlinearity serves to transform each observation into a Gaussian random variable. In the case of zero-mean Gaussian noise, $-d\ln p_N(x)/dx = x/\sigma^2$. Interestingly, this Gaussian case is the *only* instance where the "non-linearity" is indeed linear and yields a Gaussian output. Now consider the case where the noise is Laplacian $\left(p_N(N) = \frac{1}{\sqrt{2\sigma^2}}\exp\{-|N|/\sqrt{\sigma^2/2}\}\right)$. The detector's front-end transformation is $-d\ln p_N(x)/dx = \text{sign}[x]/\sqrt{\sigma^2/2}$. This non-linearity is known as an infinite clipper and the output consists *only* of the values $\pm 1/\sqrt{\sigma^2/2}$, not a very Gaussian quantity.

Robust Small-Signal Detection

These results can be modified to derive a robust detector that accommodates small signals. To make this detector insensitive to deviations from the nominal noise probability density, the density should be replaced by the worst-case density consistent with the maximum allowed deviations. For unimodal, symmetric densities, the worst-case density equals that of the nominal within a region centered about the origin while being proportional to an exponential for values outside that region.

$$p_N^w(N) = \begin{cases} (1-\varepsilon)p_N^o(N) & |N| < N' \\ (1-\varepsilon)p_N^o(N')\exp\{-a(|N|-N')\} & |N| > N' \end{cases}$$

The parameter a controls the rate of decrease of the exponential tail; it is specified by requiring that the derivative of the density's logarithm be continuous at $N = N'$. Thus, $a = -d\ln p_N^o(N)/dN$ for $N = N'$. For the worst-case density to be a probability density, the boundary value N' and the deviation parameter ε must be such that its integral is unity. As ε is fixed, this requirement reduces to a specification of N'.

$$\int_{-N'}^{N'} p_N^o(N)\,dN + \frac{2}{a}p_N^o(N') = \frac{1}{1-\varepsilon}$$

For the case where the nominal density is Gaussian, $a = N'/\sigma^2$ and this equation becomes

$$1 - 2Q\left(\frac{N'}{\sigma}\right) + \frac{\sigma}{N'}\sqrt{\frac{2}{\pi}}e^{-\frac{N'^2}{2\sigma^2}} = \frac{1}{1-\varepsilon}.$$

The non-linearity required by the small-signal detection strategy thus equals the negative derivative of the logarithm of the worst-case density within the boundaries while equaling a constant outside that region.

$$\boxed{f[X(l)] = \begin{cases} -\left.\frac{d\ln p_N^o(N)}{dN}\right|_{N=X(l)} & |X(l)| < N' \\ a\cdot\text{sign}[X(l)] & |X(l)| > N' \end{cases}}$$

The robust detector for small signals thus consists of a clipper followed by a matched filter for each signal, the results of which are compared with a threshold.

5.6.8 Non-Parametric Detection

In situations when no nominal density can be reasonably assigned or when the possible extent of deviations from the nominal cannot be assessed, non-parametric detection theory can rise to the occasion [38, 59]. In this

framework first explored in §5.6.3 {200}, little is assumed about the form of the noise density. Assume that model $\mathscr{M}_0$ corresponds to the noise-only situation and $\mathscr{M}_1$ to the presence of a signal. Moreover, assume that the noise density has zero median: any noise value is equally likely to be positive or negative. This assumption does not necessarily demand that the density be symmetric about the origin, but such densities do have zero median. Given these assumptions, the formalism of non-parametric model evaluation yields the sign test as the best decision rule. As described in the simpler model evaluation context, $\mathscr{M}_1$ had constant, positive mean for each observation. Signal values are usually unequal and change sign; we must extend the sign test to this more realistic situation. Noting that the statistic of the sign test did not depend on the value of the mean but on its sign, the sign of each observation should be "matched" with the sign of each signal value. A kind of matched filter results, where $\text{sign}[X(l)]$ is match-filtered with $\text{sign}[s(l)]$.

$$\boxed{\sum_{l=0}^{L-1} \mathrm{u}\big(\text{sign}[X(l)]\,\text{sign}[s(l)]\big) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma} \tag{5.16}$$

The quantity $\mathrm{u}(\cdot)$ is the unit-step function; the sum counts the times when the signal and the observation signs matched.

To find the threshold, the ubiquitous Central Limit Theorem can be invoked. Under $\mathscr{M}_0$, the expected value of summation is $L/2$ and its variance $L/4$. The false alarm probability for a given value of γ is therefore *approximately* given by

$$P_F = Q\left(\frac{\gamma - L/2}{\sqrt{L/4}}\right)$$

and the threshold easily found. We find the probability of detection with similar techniques, assuming a Gaussian density describes the distribution of the sum when a signal is present. Letting P_l denote the probability the observation and the signal value agree in sign at the l^{th} sample, the sum's expected value is $\sum_l P_l$ and its variance $\sum_l P_l(1-P_l)$. Using the Central Limit Theorem approximation, the probability of detection is given by

$$P_D = Q\left(\frac{\gamma - \sum_l P_l}{\sqrt{\sum_l P_l(1-P_l)}}\right).$$

For a symmetric as well as zero-median density for the noise amplitude, this probability is given by $P_l = \int_{-|s(l)|}^{\infty} p_N(N)\,dN$. If Gaussian noise were present, this probability would be $1 - Q\big(|s(l)|/\sigma\big)$ and for Laplacian noise $1 - \frac{1}{2}\exp\big\{-|s(l)|/\sqrt{\sigma^2/2}\big\}$.

The non-parametric detector expressed by Eq. (5.16) has many attractive properties for signal processing applications. First, the detector does not require knowledge of the amplitude of the signal. In addition, note that the false-alarm probability does *not* depend on the variance of the noise; the sign detector is therefore *CFAR*. Another property of the sign detector is its robustness: we have implicitly assumed that the noise values have *the* worst-case probability density—the Laplacian. A more practical property is the one bit of precision required by the quantities used in the computation of the sufficient statistic: each observation is passed through an infinite clipper (a one-bit quantizer) and matched (ANDed) with a one bit representation of the signal. A less desirable property is the dependence of the sign detector's performance on the signal waveform. A signal having a few dominant peak values may be less frequently detected than an equal energy one having a more constant envelope. As the following example demonstrates, the loss in performance compared to a detector specially tailored to the signal and noise properties can be small.

Example

Let the signal to be detected be a sinusoid having, for simplicity, an integer number of cycles within the L observations $[s(l) = \sqrt{2E/L}\sin(2\pi f_o l)]$. Setting a false alarm probability of 0.01 results in a threshold value of $\gamma = 1.16\sqrt{L} + L/2$. Fig. 5.13 {187} depicts the probability of detection for various

signal-to-noise ratios when the case $f_o = 0.1$ and $L = 100$. The loss in performance compared to the matched filter is small—approximately 3 dB!

5.6.9 Type-based detection

Perhaps the ultimate non-parametric detector makes *no* assumptions about the observations' probability distribution under either model. Here, we assume that data representative of each model are available to train the detection algorithm. One approach uses artificial neural networks, which are difficult to analyze in terms of both optimality and performance. When the observations are discrete-valued, a provably optimal detection algorithm [41] can be derived using the theory of types {128}.

For a two-model evaluation problem, let $\widetilde{\mathbf{X}}$ (length $\widetilde{L}$) denote training data representative of some unknown probability distribution P. We assume that the data have statistically independent components. To derive a non-parametric detector, form a generalized likelihood ratio to distinguish whether a set of observations $\mathbf{X}$ (length L) has the same distribution as the training data or a different one Q.

$$\log\Lambda(\mathbf{X}) = \log \frac{\max_{P,Q} P(\mathbf{X})Q(\widetilde{\mathbf{X}})}{\max_P P(\mathbf{X})P(\widetilde{\mathbf{X}})}$$

Because a type is the maximum likelihood estimate of the probability distribution {129}, we simply substitute types for the training data and observations probability distributions into the likelihood ratio. The probability of a set of observations having a probability distribution identical to its type equals $e^{-L\mathscr{H}(\widehat{P})}$. Thus, the log likelihood ratio becomes

$$\log\Lambda(\mathbf{X}) = \log \frac{e^{-L\mathscr{H}(\widehat{P}_X)} e^{-\widetilde{L}\mathscr{H}(\widehat{P}_{\tilde{X}})}}{e^{-(L+\widetilde{L})\mathscr{H}(\widehat{P}_{X,\tilde{X}})}} .$$

The denominator term means that the training and observed data are lumped together to form a type. This type equals the linear combination of the types for the training and observed data weighted by their relative lengths.

$$\widehat{P}_{X,\tilde{X}} = \frac{L\widehat{P}_X + \widetilde{L}\widehat{P}_{\tilde{X}}}{L+\widetilde{L}}$$

Returning to the log likelihood ratio, we have that

$$\log\Lambda(\mathbf{X}) = -L\mathscr{H}(\widehat{P}_X) - \widetilde{L}\mathscr{H}(\widehat{P}_{\tilde{X}}) + (L+\widetilde{L})\mathscr{H}\left(\frac{L\widehat{P}_X + \widetilde{L}\widehat{P}_{\tilde{X}}}{L+\widetilde{L}}\right)$$

Note that the last term equals

$$(L+\widetilde{L})\mathscr{H}\left(\frac{L\widehat{P}_X + \widetilde{L}\widehat{P}_{\tilde{X}}}{L+\widetilde{L}}\right) = -\sum_a \left[L\widehat{P}_X(a) + \widetilde{L}\widehat{P}_{\tilde{X}}(a)\right] \log \frac{L\widehat{P}_X(a) + \widetilde{L}\widehat{P}_{\tilde{X}}(a)}{L+\widetilde{L}} ,$$

which means it can be combined with the other terms to yield a simple expression for the log likelihood ratio in terms of the Kullback-Leibler distance.

$$\boxed{\log\Lambda(\mathbf{X}) = L\mathscr{D}(\widehat{P}_X \| \widehat{P}_{X,\tilde{X}}) + \widetilde{L}\mathscr{D}(\widehat{P}_{\tilde{X}} \| \widehat{P}_{X,\tilde{X}})}$$

When the training data and the observed data are drawn from the same distribution, the Kullback-Leibler distances will be small. When the distributions differ, the distances will be larger. Defining $\mathscr{M}_0$ to be the model that the training data and observations have the same distribution and $\mathscr{M}_1$ that they don't, Gutman [41] showed that when we use the decision rule

$$\frac{1}{L}\log\Lambda(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma ,$$

its false-alarm probability has an exponential rate at least as large as the threshold and the miss probability is the smallest among *all* decision rules based on training data.

$$\lim_{L\to\infty}\frac{1}{L}\log P_F \le -\gamma \quad \text{and} \quad P_M \text{ minimum}$$

We can extend these results to the K-model case if we have training data $\widetilde{\mathbf{X}}_i$ (each of length $\widetilde{L}$) that represent model $\mathscr{M}_i$, $i=0,\ldots,K-1$. Given observed data $\mathbf{X}$ (length L), we calculate the log likelihood function given above for each model to determine whether the observations closely resemble the tested training data or not. More precisely, define the sufficient statistics Υ_i according to

$$\Upsilon_i = \mathscr{D}(\widehat{P}_X \| \widehat{P}_{X,\tilde{X}_i}) + \frac{\widetilde{L}}{L}\mathscr{D}(\widehat{P}_{\tilde{X}} \| \widehat{P}_{X,\tilde{X}_i}) - \gamma\,.$$

Ideally, this statistic would be negative for one of the training sets (matching it) and positive for all of the others (not matching them). However, we could also have the observation matching more than one training set. In all such cases, we define a *rejection region* $\Re_?$ similar to what we defined in sequential model evaluation 5.2 {165}. Thus, we define the i^{th} decision region $\Re_i$ according to $\Upsilon_i < 0$ and $\Upsilon_j > 0$, $j \neq i$ and the rejection region as the complement of $\bigcup_{i=0}^{K-1}\Re_i$. Note that all decision regions depend on the value of γ, a number we must choose. Regardless of the value chosen, the probability of confusing models—choosing some model other than the true one—has an exponential rate that is at least γ for all models. Because of the presence of a rejection region, another kind of "error" is to not choose any model. This decision rule is optimal in the sense that no other training-data-based decision rule has a smaller rejection region that the type-based one.

Because it controls the exponential rate of confusing models, we would like γ to be as large as possible. However, the rejection region grows as γ increases; choosing too large a value could make virtually all decisions rejections. What we want to ensure is that $\lim_{L\to\infty}\Pr[\Re_?|\mathscr{M}_i]=0$. Obtaining this behavior requires that $\lim_{L\to\infty}\frac{\widetilde{L}}{L}>0$: As the length of the observations increases, so must the size of the training set. In summary,

$$\text{If } \lim_{L\to\infty}\frac{\widetilde{L}}{L}=0 \quad \text{or} \quad \gamma>\gamma_0 \implies \lim_{L\to\infty}\Pr[\Re_?|\mathscr{M}_i]=1, \text{for some } i$$

$$\text{If } \lim_{L\to\infty}\frac{\widetilde{L}}{L}>0 \quad \text{and} \quad \gamma<\gamma_0 \implies \lim_{L\to\infty}\frac{1}{L}\Pr[\Re_?|\mathscr{M}_i] \le -\beta < 0, \forall i$$

The critical value γ_0 depends on the true distributions underlying the models. The exponential rate of the rejection probability β also depends on the true distributions. These results mean that if sufficient training data are available and the decision threshold is not too large, then we can perform optimal detection based entirely on data! As the number of observations increases (and the amount of training data as well), the critical threshold γ_0 becomes the Kullback-Leibler distance between the unknown models. In other words, the type-based detector becomes optimal!

Problems

5.1 Consider the following two-model evaluation problem [89: Prob. 2.2.1].

$$\mathscr{M}_0 : X = N$$
$$\mathscr{M}_1 : X = s + N,$$

where s and N are statistically independent, positively valued, random variables having the densities

$$p_s(s) = ae^{-as} \quad \text{and} \quad p_N(N) = be^{-bN}.$$

(a) Prove that the likelihood ratio test reduces to

$$X \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma$$

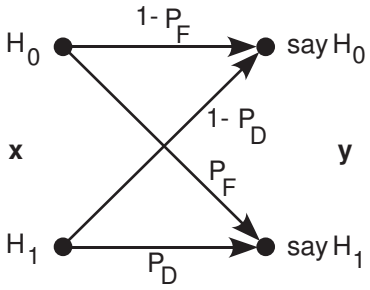


Figure 5.24: The two-model testing problem can be abstractly described as a communication channel where the inputs are the models and the outputs are the decisions. The transition probabilities are related to the false-alarm (P_F) and detection (P_D) probabilities.

(b) Find γ for the minimum probability of error test as a function of the *a priori* probabilities.

(c) Now assume that we need a Neyman-Pearson test. Find γ as a function P_F, the false-alarm probability.

5.2 Two models describe different equi-variance statistical models for the observations [89: Prob. 2.2.11].

$$\mathscr{M}_0 : \; p_X(X) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|X|}$$

$$\mathscr{M}_1 : \; p_X(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}X^2}$$

(a) Find the likelihood ratio.

(b) Compute the decision regions for various values of the threshold in the likelihood ratio test.

(c) Assuming these two densities are equally likely, find the probability of making an error in distinguishing between them.

5.3 Cautious Decision Making

Wanting to be very cautious before making a decision, an ELEC 531 student wants to explicitly allow no decision to be made if the data don't warrant a "firm" decision. The Bayes cost criterion is used to derive the cautious detector. Let the cost of a wrong decision be $C_{\text{wd}} > 0$, the cost of making no decision be $C_? > 0$ and the cost of making a correct decision be zero. Two signal models are possible and they have *a priori* probabilities π_0 and π_1.

(a) Derive the detector that minimizes the average Bayes cost, showing that it is a likelihood ratio detector.

(b) For what choices of C_{wd} and $C_?$ is the decision rule well-defined?

(c) Let the observations consist of L samples, with model 0 corresponding to white Gaussian noise and model 1 corresponding to a known signal in additive white Gaussian noise. Find the decision rule in terms of the sufficient statistic.

5.4 A hypothesis testing criterion radically different from those discussed in §5.1.1 and §5.1.2 is *minimum equivocation*. In this information theoretic approach, the two-model testing problem is modeled as a digital channel (Fig. 5.24). The channel's inputs, generically represented by the $\mathbf{x}$, are the models and the channel's outputs, denoted by $\mathbf{y}$, are the decisions.

The quality of such information theoretic channels is quantified by the *mutual information* $I(\mathbf{x};\mathbf{y})$ defined to be difference between the entropy of the inputs and the *equivocation* [23: §2.3, 2.4].

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y})$$

where

$$H(\mathbf{x}) = -\sum_i \pi_i \log \pi_i$$

$$H(\mathbf{x}|\mathbf{y}) = -\sum_{i,j} P(x_i, y_j) \log P(x_i|y_j) = -\sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(y_j)}$$

Here, π_i denotes the *a priori* probabilities, $P(y_j)$ the output probabilities (i.e., the decision probabilities), and $P(x_i, y_j)$ the joint probability of input x_i resulting in the decision y_j. For example, $P(x_0, y_0) = \pi_0(1-P_F)$ and $P(y_0) = \pi_0(1-P_F) + \pi_1(1-P_D)$.

In most information theoretic problems, the cross-over probabilities that characterize the channel are fixed and we try to maximize mutual information by varying the *a priori* probabilities. In this problem, the *a priori* probabilities are fixed and we want to find the optimal "channel." Let's take a Neymann-Pearson approach: fix P_F and try to maximize $I(\mathbf{x};\mathbf{y})$ with respect to P_D. Since the entropy $H(\mathbf{x})$ is fixed in this approach, we just need to minimize the equivocation $H(\mathbf{x}|\mathbf{y})$ to maximize mutual information.

(a) Find all of the values of P_D that correspond to stationary points of the mutual information for a fixed values of P_F, π_0 and π_1.

(b) Show that the stationary point(s) you found correspond to *minima*, not maxima, of mutual information.

(c) What situation(s) characterize the occurrence of these minima?

(d) What decision process maximizes mutual information?

5.5 Detection and Estimation Working Together

Detectors are frequently used to determine if a signal is even present before applying estimators to tease out the signal. The same, but unknown signal having duration L may or may not be present in additive white Gaussian noise during the i^{th} observation interval, $i = 1, \ldots$.

$$\mathscr{M}_0 \colon \mathbf{X}_i = \mathbf{N}_i$$
$$\mathscr{M}_1 \colon \mathbf{X}_i = \mathbf{s} + \mathbf{N}_i$$

π_0, π_1 denote the *a priori* probabilities. Once M intervals have been determined by the front-end detector to contain a signal, we apply the maximum likelihood estimator to measure the signal.

(a) What is the maximum likelihood signal estimate?

(b) What is the front-end detector's algorithm?

(c) Even if we use an optimal front-end detector, it can make errors, saying a signal is present when it isn't. What is the mean-squared error of the combined detector-estimator in terms of the detector's detection and false-alarm probabilities?

5.6 Non-Gaussian statistical models sometimes yield surprising results in comparison to Gaussian ones. Consider the following hypothesis testing problem where the observations have a Laplacian probability distribution.

$$\mathscr{M}_0 \colon p_X(X) = \frac{1}{2} e^{-|X+m|}$$
$$\mathscr{M}_1 \colon p_X(X) = \frac{1}{2} e^{-|X-m|}$$

(a) Find the sufficient statistic for the optimal decision rule.

(b) What decision rule guarantees that the miss probability will be less than 0.1?

5.7 Developing a Neyman-Pearson decision rule for more than two models has not been detailed because a mathematical quandry arises. The issue is that we have several performance probabilities we want to optimize. In essence, we are optimizing a vector of performance probabilities, which requires us to specify a norm. Many norms can be chosen; we select one in this problem.

Assume K distinct models are required to account for the observations. We seek to maximize the *sum* of the probabilities of correctly announcing $\mathscr{M}_i$, $i = 1, \ldots, K$. This choice amounts to maximizing the L^1 norm of the detection probabilities. We constrain the probability of announcing $\mathscr{M}_i$ when model $\mathscr{M}_0$ was indeed true to not exceed a specified value.

(a) Formulate the optimization problem that simultaneously maximizes $\sum_{i'} \Pr[\text{say } \mathscr{M}_i | \mathscr{M}_i]$ under the constraint $\Pr[\text{say } \mathscr{M}_i | \mathscr{M}_0] \leq \alpha_i$. Find the solution using Lagrange multipliers.

(b) Can you find the Lagrange multipliers?

(c) Can your solution can be expressed as choosing the largest of the sufficient statistics $\Upsilon_i(\mathbf{X}) + C_i$?

5.8 Pattern recognition relies heavily on ideas derived from the principles of statistical model testing. Measurements are made of a test object and these are compared with those of "standard" objects to determine which the test object most closely resembles. Assume that the measurement vector $\mathbf{X}$ is jointly Gaussian with mean $\mathbf{m}_i$ $(i = 1, \ldots, K)$ and covariance matrix $\sigma^2\mathbf{I}$ (i.e., statistically independent components). Thus, there are K possible objects, each having an "ideal" measurement vector $\mathbf{m}_i$ and probability π_i of being present.

(a) How is the minimum probability of error choice of object determined from the observation of $\mathbf{X}$?

(b) Assuming that only two equally likely objects are possible ($K = 2$), what is the probability of error of your decision rule?

(c) The expense of making measurements is always a practical consideration. Assuming each measurement costs the same to perform, how would you determine the effectiveness of a measurement vector's component?

5.9 Define y to be

$$y = \sum_{k=0}^{L} x_k$$

where the x_k are statistically independent random variables, each having a Gaussian density $\mathcal{N}(0, \sigma^2)$. The number L of variables in the sum is a random variable with a Poisson distribution.

$$\Pr[L = l] = \frac{\lambda^l}{l!} e^{-\lambda}, \quad l = 0, 1, \ldots$$

Based upon the observation of y, we want to decide whether $L \leq 1$ or $L > 1$. Write an expression for the minimum P_e likelihood ratio test.

5.10 One observation of the random variable X is obtained. This random variable is either uniformly distributed between -1 and $+1$ or expressed as the sum of statistically independent random variables, each of which is also uniformly distributed between -1 and $+1$.

(a) Suppose there are two terms in the aforementioned sum. Assuming that the two models are equally likely, find the minimum probability of error decision rule.

(b) Compute the resulting probability of error of your decision rule.

(c) Show that the decision rule found in part (a) applies no matter how many terms are assumed present in the sum.

5.11 The observed random variable X has a Gaussian density on each of five models.

$$p_{X|\mathscr{M}_i}(X|\mathscr{M}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(X - m_i)^2}{2\sigma^2}\right\}, \quad i = 1, 2, \ldots, 5$$

where $m_1 = -2m$, $m_2 = -m$, $m_3 = 0$, $m_4 = +m$, and $m_5 = +2m$. The models are equally likely and the criterion of the test is to minimize P_e.

(a) Draw the decision regions on the X-axis.

(b) Compute the probability of error.

(c) Let $\sigma = 1$. Sketch accurately P_e as a function of m.

5.12 The goal is to choose which of the following four models is true upon the reception of the three-dimensional vector $\mathbf{X}$ [89: Prob. 2.6.6].

$$\mathscr{M}_0 : \mathbf{X} = \mathbf{m}_0 + \mathbf{N}$$
$$\mathscr{M}_1 : \mathbf{X} = \mathbf{m}_1 + \mathbf{N}$$
$$\mathscr{M}_2 : \mathbf{X} = \mathbf{m}_2 + \mathbf{N}$$
$$\mathscr{M}_3 : \mathbf{X} = \mathbf{m}_3 + \mathbf{N}$$

where

$$\mathbf{m}_0 = \begin{bmatrix} a \\ 0 \\ b \end{bmatrix}, \quad \mathbf{m}_1 = \begin{bmatrix} 0 \\ a \\ b \end{bmatrix}, \quad \mathbf{m}_2 = \begin{bmatrix} -a \\ 0 \\ b \end{bmatrix}, \quad \mathbf{m}_3 = \begin{bmatrix} 0 \\ -a \\ b \end{bmatrix}.$$

The noise vector $\mathbf{N}$ is a Gaussian random vector having statistically independent, identically distributed components, each of which has zero mean and variance σ^2. We have L independent observations of the received vector $\mathbf{X}$.

(a) Assuming equally likely models, find the minimum P_e decision rule.

(b) Calculate the resulting error probability.

(c) Show that neither the decision rule nor the probability of error do not depend on b. Intuitively, why is this fact true?

5.13 Discrete Estimation

Estimation theory focuses on deriving effective (minimum error) techniques for determining the value of *continuous-valued* quantities. When the quantity is discrete-valued (integer-valued, for example), the usual approaches don't work well since they usually produce estimates not in the set of known values. This problem explores applying decision-theoretic approaches to yield a framework for discrete estimation.

Let's explore a specific example. Let a sequence of statistically independent, identically distributed observations be Gaussian having mean m and variance σ^2. The mean m can only assume the values -1, 0, and 1, and these are equally likely. The mean, whatever its value, is constant throughout the L observations.

(a) What is the minimum probability of error decision rule? What is the resulting probability of error?

(b) What is the MAP estimate of the mean?

(c) The problem with using the detection approach of part (a) is that the probability of error is not a standard error metric in estimation theory. Suppose we want to find the minimum mean-squared error, discrete-valued estimate. Show that by defining an appropriate Bayes cost function, you can create a detection problem that minimizes the mean-squared error. What is the Bayes cost function that works?

(d) Find the minimum mean-squared error estimate using the minimum Bayes cost detector.

(e) What is the resulting mean-squared error?

5.14 Diversity Communication

In diversity signaling, one of two equally likely signal groups is transmitted, with each member $\mathbf{s}_m$ of the group $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M\}$ sent through one of M parallel channels simultaneously. The receiver has access to all channels and can use them to make a decision as to which signal group was transmitted. The received vector (dimension L) emerging from the m^{th} channel has the form

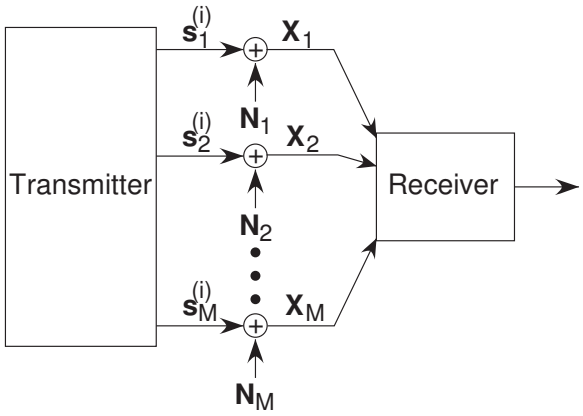
$$\mathbf{X}_m = \mathbf{s}_m^{(i)} + \mathbf{N}_m, \quad i = 0, 1$$

The noise is colored, Gaussian, and independent from channel to channel. The statistical properties of the noise are known and they are the same in each channel.

(a) What is the optimal diversity receiver?

In an endeavor to make communications secure, bright engineers decide to modify the usual diversity signaling scenario. Rather than send signals down each channel, only one channel will be used for each transmission, with the chosen channel selected randomly from transmission to transmission. The identity of the used channel is kept secret, even from the receiver. In essence, each group now mostly consists of zero-valued signals save for the one $\mathbf{s}_m$ the transmitter chose. The receiver has access to all channels and must determine which of the signal groups was sent. In this scenario, the channel actually used contains no information about the signal group.

(b) How would the receiver determine which channel was used?



(c) What is the optimal decision rule?

5.15 Detecting Random Signals

In the framework for Wiener filters, the signal as well as the noise is considered a random process. What happens when we make a similar assumption in the context of a detection problem? Assume we have two models for the observations, one without a signal and one with.

$$\mathscr{M}_0 : \mathbf{X} = \mathbf{N}$$
$$\mathscr{M}_1 : \mathbf{X} = \mathbf{S} + \mathbf{N}$$

Here, both the signal and the noise are zero-mean Gaussian random vectors having covariance matrices $\mathbf{K}_S$ and $\mathbf{K}_N$, respectively.

(a) Assuming the two models are equally likely, what decision rule minimizes the probability of a decision error?

(b) Suppose the noise is white with variance σ_N^2 and the signal is also white, but with an unknown variance. Now what is the decision rule that minimizes the probability of a decision error?

5.16 Keeping Communications Secret

In a digital communication scheme, the signal set consists of sinusoids having different frequencies. Which signal is sent corresponds to the data bit; data bits are sent one after the other, the signal chosen according to the bit. Assume the data bits form an IID sequence of equally likely Bernoulli (binary-valued) random variables. Since the receiver is implemented digitally, the received signal is sampled, making the received signal to have one of two forms.

$$\mathscr{M}_0 : X_\ell = \sqrt{\frac{2E}{L}} \sin 2\pi f_0 \ell + N_\ell \quad 0 \le \ell \le L-1$$
$$\mathscr{M}_1 : X_\ell = \sqrt{\frac{2E}{L}} \sin 2\pi f_1 \ell + N_\ell \quad 0 \le \ell \le L-1$$

E is the received signal energy expended over L samples and $f_0 < f_1$. The frequencies are chosen to have an integer number of cycles within the observation interval. The noise N_ℓ is white and Gaussian.

(a) What is the optimal (minimum probability of error) receiver for this digital communication scheme?

(b) This approach is not very secretive: all a snooper would need to do is listen to the transmissions and look for two frequencies. Assuming the snooper knew the frequency band that contained f_0 and f_1 ($f_L \le f_0 < f_1 \le f_U$), design a snooping algorithm.

(c) In an attempt to confound the snooper, the transmitter will randomly switch frequencies. Instead of using frequency f_0, he will use either $f_0^{(1)}$ or $f_0^{(2)}$; instead of f_1, either $f_1^{(1)}$ or $f_1^{(2)}$. All of these frequencies lie in the same frequency band. Does this approach confound the snooper? In other words, will the snooper be less successful in listening in?

5.17 Color Image Detection

We acquire color images, which are formatted as a vector having three components that express the component images in red, green and blue.

$$\mathbf{X} = \begin{bmatrix} \mathbf{r} \\ \mathbf{g} \\ \mathbf{b} \end{bmatrix}$$

Each color component in this expression is derived from the image (an image amounts to a matrix) by vectorizing it, thereby representing each color component as a column vector. Each color component—**r**, **g**, **b**—is described as a Gaussian random vector having a positive-valued mean and a covariance. The covariance matrices for the components equal each other; furthermore, the components are correlated with each other, each pair having the same cross-covariance matrix.

(a) Suppose we want to determine if one of two nominal images is the observed image. The two images have the same covariance matrices but different means. Assuming the images are equally likely, what is the minimum probability of error decision rule?

(b) How well will the decision rule work?

(c) I.M. Cheap wants to simplify the decision system by using grayscale images. The grayscale version of $\mathbf{X}$ equals $\mathbf{r}+\mathbf{g}+\mathbf{b}$. What is the optimal decision rule that uses grayscale images?

(d) Do you expect the grayscale system to perform better or worse than the color-based one? Why?

5.18 To gain some appreciation of some of the issues in implementing a detector, this problem asks you to program (preferably in MATLAB) a simple detector and numerically compare its performance with theoretical predictions. Let the observations consist of a signal contained in additive Gaussian white noise.

$$\mathscr{M}_0 : X(l) = N(l), \quad l = 0,\ldots,L-1$$
$$\mathscr{M}_1 : X(l) = A\sin(2\pi l/L) + N(l), \quad l = 0,\ldots,L-1$$

The variance of each noise value equals σ^2.

(a) What is the theoretical false-alarm probability of the minimum P_e detector when the hypotheses are equally likely?

(b) Write a MATLAB program that estimates the false-alarm probability. How many simulation trials are needed to accurately estimate the false-alarm probability? Choose values for A and σ^2 that will result in values for P_F of 0.1 and 0.01. Estimate the false-alarm probability and compare with the theoretical value in each case.

5.19 The Kullback-Leibler and Chernoff distances can be related to the Fisher information matrix defined in (4.11) {70}. Let $p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})$ be a probability density that depends on the parameter vector $\boldsymbol{\theta}$. In this problem, use the short-hand notation $p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) = p(\mathbf{X};\boldsymbol{\theta})$. We want to consider the distance between probability densities that differ by a perturbation $\boldsymbol{\delta\theta}$ in their parameter vectors.

(a) Show that Ali-Silvey distances, defined in appendix C {247}, have the property

$$\left.\frac{\partial^2 d(p(\mathbf{X};\boldsymbol{\theta}), p(\mathbf{X};\boldsymbol{\theta}_0))}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \propto [\mathbf{F}(\boldsymbol{\theta}_0)]_{i,j}$$

The Ali-Silvey distance family consists of distance measures defined by

$$d(p(\mathbf{X};\boldsymbol{\theta}), p(\mathbf{X};\boldsymbol{\theta}_0)) = f\left(\int p(\mathbf{X};\boldsymbol{\theta}) c\big(\Lambda(\mathbf{X})\big)\, d\mathbf{X}\right), \ \Lambda(\mathbf{X}) = \frac{p(\mathbf{X};\boldsymbol{\theta}_0)}{p(\mathbf{X};\boldsymbol{\theta})}$$

where $f(\cdot)$ is non-decreasing and $c(\cdot)$ is convex. What is the constant of proportionality?

(b) Using this result, show that

$$\mathscr{D}(p(\mathbf{X};\boldsymbol{\theta}_0+\boldsymbol{\delta\theta})\|p(\mathbf{X};\boldsymbol{\theta}_0)) \propto (\boldsymbol{\delta\theta})'\mathbf{F}(\boldsymbol{\theta}_0)\boldsymbol{\delta\theta} \quad \text{for small } \boldsymbol{\delta\theta}.$$

What is the constant of proportionality?

(c) Despite the Chernoff distance not belonging to the Ali-Silvey class, it has the property studied in part (a). Show that this statement is true and find the constant of proportionality.

5.20 Calculate the Kullback-Leibler distance between the following pairs of densities.

(a) Jointly Gaussian random vectors having the same covariance matrix but dissimilar mean vectors.

(b) Two Poisson random variables having average rates λ_0 and λ_1. In this example, the observation time T plays the role of the number of observations.

(c) Two sequences of statistically independent Laplacian random variables having the same variance but different means.

(d) Plot the Kullback-Leibler distances for the Laplacian case and for the Gaussian case of statistically independent random variables. Let the difference between the means be m and set the variance equal to σ^2 in each case. Plot the distances as a function of m/σ.

(e) Use these results to find the Fisher information for the mean parameter m for the Gaussian and Laplacian cases.

5.21 Insights into certain *detection* problems can be gained by examining the Kullback-Leibler distance and the properties of Fisher information. We begin by first showing that the Gaussian distribution has the smallest Fisher information for the mean parameter for all differentiable distributions having the same variance.

(a) Show that if $f(t)$ and $g(t)$ are linear functions of t and $g(t)$ is positive for $0 < t < 1$, then the ratio $[f(t)]^2/g(t)$ is convex over this interval.

(b) Use this property to show that the Fisher information is a convex function of the probability density.

(c) Define $p_t(x;\theta) = (1-t)p_0(x;\theta) + t\,p_1(x;\theta)$, $0 \le t \le 1$, where $p_0(x;\theta), p_1(x;\theta) \in \mathscr{P}$, a class of densities having variance one. Show that this set is convex.

(d) Because of the Fisher information's convexity, a given distribution $p_0(x|\theta) \in \mathscr{P}$ minimizes the Fisher information if and only if $\frac{d}{dt}\mathbf{F}_t \ge 0$ at $t = 0$ for all $p_1(x|\theta) \in \mathscr{P}$. Let the parameter be the expected value of all densities in $\mathscr{P}$. By using Lagrange multipliers to impose the constant variance constraint on all densities in the class, show that the Gaussian uniquely minimizes Fisher information.

(e) What does this result suggest about the performance probabilities for problems wherein the models differ in mean?

5.22 Find the Chernoff distance between the following distributions.

(a) Two Gaussian distributions having the same variance but different means.

(b) Two Poisson distributions having differing parameter values.

5.23 Let's explore how well Stein's Lemma predicts optimal detector performance probabilities. Consider the two-model detection problem wherein L statistically independent, identically distributed Gaussian random variables are observed. Under $\mathscr{M}_0$, the mean is zero and the variance one; under $\mathscr{M}_1$, the mean is one and the variance one.

(a) Find an expression for the false-alarm probability P_F when the miss probability is constrained to be less than α.

(b) Find the Kullback-Leibler distance corresponding to the false-alarm probability's exponent.

(c) Plot the exact error for values of $\alpha = 0.1$ and 0.01 as a function of the number of observations. Plot on the same axes the result predicted by Stein's Lemma.

5.24 We observe a Gaussian random variable X. This random variable has zero mean under model $\mathscr{M}_0$ and mean m under $\mathscr{M}_1$. The variance of X in either instance is σ^2. The models are equally likely.

(a) What is an expression for the probability of error for the minimum P_e test when one observation of X is made?

(b) Assume one can perform statistically independent observations of X. Construct a sequential decision rule which results in a P_e equal to one-half of that found in part (a).

(c) What is the sufficient statistic in part (b) and sketch how the thresholds for this statistic vary with the number of trials. Assume that $m = 10$ and that $\sigma = 1$. What is the expected number of trials for the sequential test to terminate?

5.25 Sequential Null-Hypothesis Testing

You are provided with a sequence of statistically independent, identically distributed observations. Your boss wants you to determine if the data conform to a specific probabilistic model or not. Because making a decision quickly is important, you decide to use a sequential approach.

Because the standard sequential approach places criteria on P_F, the probability of saying the model applies when it doesn't, and P_D, the probability you say the model applies when it does, an issue arises immediately: P_F cannot be defined because no explicit alternative model exists. Instead, you decide to place constraints on P_D and P_M, the probability you say the model doesn't apply when it does.

(a) Are there restrictions on the values of $P_D = \alpha$ and $P_M = \beta$ that can be used as specifications?

(b) For specified values of α and β, what is the resulting sequential decision rule?

(c) Explicitly derive the sequential test when the nominal model for each observation is a zero-mean Gaussian random variable having unit variance.

5.26 Drug Testing

In testing drugs, the variability among patients makes judging effectiveness difficult, but not impossible. The number of people N a drug cures has a geometric probability distribution.

$$\Pr[N = n] = (1-a)a^n,\ n = 0, 1, \ldots$$

You perform a drug trial over a very large population (large enough so that the approximation of the geometric probability distribution remains valid). Either the drug is ineffective, in which case the distribution's parameter equals a_0, or is effective and the parameter equals a_*, $a_* > a_0$. The *a priori* probability that the drug will be effective is π_*.

(a) Construct the minimum probability of error test that decides drug effectiveness.

(b) What is the probability of error the test will achieve?

(c) Now suppose the drug trial is repeated in several countries, each of which has a large population. Because conducting such tests is expensive, you want a test that will reach a conclusion as quickly as possible. Find a test that will achieve false-positive and false-negative rates of α and $1-\beta$, respectively, as quickly as possible.

5.27 The optimum reception of binary information can be viewed as a model testing problem. Here, equally-likely binary data (a "zero" or a "one") is transmitted through a binary symmetric channel (Figure 5.25). The indicated parameters denote the probabilities of receiving a binary digit given that a particular digit was sent. Assume that $\varepsilon = 0.1$.

(a) Assuming a single transmission for each digit, what is the minimum probability of error receiver and what is the resulting probability of error?

(b) One method of improving the probability of error is to repeat the digit to be transmitted L times. This transmission scheme is equivalent to the so-called *repetition code*. The receiver uses all of the received L digits to decide what digit was actually sent. Assume that the results of each transmission are statistically independent of all others. Construct the minimum probability of error receiver and find an expression for P_e in terms of L.

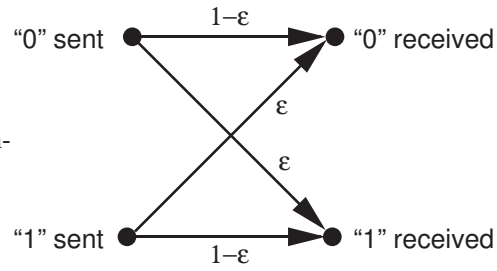


Figure 5.25: A binary symmetric digital communications channel.

(c) Assume that we desire the probability of error to be 10^{-6}. How long a repetition code is required to achieve this goal for the channel given above? Assume that the leading term in the probability of error expression found in part (b) dominates.

(d) Construct a sequential algorithm which achieves the required probability of error. Assume that the transmitter will repeat each digit until informed by the receiver that it has determined what digit was sent. What is the expected length of the repetition code in this instance?

5.28 Maybe It's There, Maybe Not...

You are provided with M examples of a length-L unknown but unit-energy signal measured in the presence of additive white Gaussian noise. The problem is that the signal may or not be present in each example: the probability the signal is present in each example is p. The noise components in each example are statistically independent of each other.

(a) Suppose for a moment that the signal is known. What procedure would you use to test whether a single example contained a signal or not?

(b) Assume you don't know p and you know the signal. What equation would need to be solved to find the maximum likelihood estimate of p?

(c) Assuming the signal is not known, find the best linear estimate of the signal assuming you know p.

5.29 Signal Amplitude Problems

The signal waveform may be known, but its amplitude A is not well specified.

$$\mathscr{M}_0 : \mathbf{X} = \mathbf{N}$$
$$\mathscr{M}_1 : \mathbf{X} = A\mathbf{s} + \mathbf{N}$$

In both models, the noise is white and Gaussian. The nominal signal has unit energy ($\|\mathbf{s}\|^2 = 1$) and the models are equally likely.

(a) In the simplest case, the amplitude A is known. What is the minimum probability of error decision rule? What is the resulting probability of error?

(b) In a more complicated situation, the amplitude equals plus or minus one ($A = \pm 1$); which is not known. Now what is the decision rule and its probability of error?

(c) Suppose in addition to not knowing the sign of the signal's amplitude, the variance of the noise is not known. How does the decision rule change?

5.30 Unknown DC Offset

We want to detect the presence or absence of a signal in noise which could be contaminated by a constant offset.

$$\mathscr{M}_0 : X(\ell) = c + N_\ell$$
$$\mathscr{M}_1 : X(\ell) = c + s(\ell) + N_\ell$$

The noise N_ℓ is white Gaussian noise. The signal $s(\ell)$ is known and has an average value over the observation interval $[0, L-1]$ of zero.

$$\sum_{\ell=0}^{L-1} s(\ell) = 0$$

The signal is as likely to be present as not.

(a) If the offset c is known, what is the minimum probability of error detector?

(b) Suppose the offset is not known. Find the minimum probability of error detector, noting how the unknown offset affects the detector.

(c) Assuming equally likely models and an unknown offset, find the smallest possible value for the probability of error.

5.31 Life-Saving Decision Rule

You have accepted the (dangerous) job of determining whether the radioactivity levels at the Chernobyl reactor are elevated or not. Because you want to stay as short a time as possible to render your professional opinion, you decide to use a sequential-like decision rule. Radioactivity is governed by Poisson probability laws, which means that the probability that n counts are observed in T seconds equals

$$\Pr[n] = \frac{(\lambda T)^n e^{-\lambda T}}{n!},$$

where λ is the radiation intensity. Safe radioactivity levels occur when $\lambda = \lambda_0$ and unsafe ones at $\lambda = \lambda_1, \lambda_1 > \lambda_0$.

(a) Construct a sequential decision rule to determine whether it is safe or not. Assume you have defined false-alarm and miss probabilities according to accepted "professional standards." According to these standards, these probabilities equal each other.

(b) What is the expected time it will take to render a decision?

5.32 Sequential tests can be used to advantage in situations where analytic difficulties obscure the problem. Consider the case where the observations either contain no signal ($\mathcal{M}_0$) or a signal whose components are randomly set to zero.

$$\mathcal{M}_0 : X_l = N_l$$

$$\mathcal{M}_1 : X_l = N_l + a_l s_l, \quad a_l = \begin{cases} 1, & \text{Prob} = p \\ 0, & \text{Prob} = 1-p \end{cases}$$

The probability that a signal value remains intact is a known quantity p and "drop-outs" are statistically independent of each other and of the noise. This kind of model is often used to describe intermittent behavior in electronic equipment.

(a) Find the likelihood ratio for the observations.

(b) Develop a sequential test that would determine whether a signal is present or not.

(c) Find a formula for the test's thresholds in terms of P_F and P_D.

(d) How does the average number of observations vary with p?

5.33 In some cases it might be wise to *not* make a decision when the data do not justify it. Thus, in addition to declaring that one of two models occurred, we might declare "no decision" when the data are indecisive. Assume you observe L statistically independent observations X_l, each of which is Gaussian and has a variance of two. Under one model the mean is zero, and under the other the mean is one. The models are equally likely to occur.

(a) Construct a hypothesis testing rule that yields a probability of no-decision no larger than some specified value α, maximizes the probabilities of making correct decisions when they are made, and makes these correct-decision probabilities equal.

(b) What is the probability of a correct decision for your rule?

5.34 You decide to flip coins with Sleazy Sam. If heads is the result of a coin flip, you win one dollar; if tails, Sam wins a dollar. However, Sam's reputation has preceded him. You suspect that the probability of tails, p, may not be $1/2$. You want to determine whether a biased coin is being used or not after observing the results of three coin tosses.

(a) You suspect that $p = 3/4$. Assuming that the probability of a biased coin equals that of an unbiased coin, how would you decide whether a biased coin is being used or not in a "good" fashion?

(b) Using your decision rule, what is the probability that your determination is incorrect?

(c) One potential flaw with your decision rule is that a specific value of p was assumed. Can a reasonable decision rule be developed without knowing p? If so, demonstrate the rule; if not, show why not.

5.35 When a patient is screened for the presence of a disease in an organ, a section of tissue is viewed under a microscope and a count of abnormal cells made. Even under healthy conditions, a small number of abnormal cells will be present. Presumably a much larger number will be present if the organ is diseased. Assume that the number L of abnormal cells in a section is geometrically distributed.

$$\Pr[L = l] = (1-\alpha)\alpha^l, l = 0, 1, \ldots$$

The parameter α of a diseased organ will be larger than that of a healthy one. The probability of a randomly selected organ being diseased is p.

(a) Assuming that the value of the parameter α is known in each situation, find the best method of deciding whether an organ is diseased.

(b) Using your method, a patient was said to have a diseased organ. In this case, what is the probability that the organ is diseased?

(c) Assume that α is known only for healthy organs. Find the disease screening method that minimizes the maximum possible value of the probability that the screening method will be in error.

5.36 How can the standard sequential test be extended to unknown parameter situations? Formulate the theory, determine the formulas for the thresholds. How would you approach finding the average number of observations?

5.37 A common situation in statistical signal processing problems is that the variance of the observations is unknown (there is no reason that noise should be nice to us!). Consider the two Gaussian model testing problem where the model differ in their means and have a common, but unknown variance.

$$\begin{array}{ll} \mathcal{M}_0: & \mathbf{X} \sim \mathcal{N}(\mathbf{0},\, \sigma^2\mathbf{I}) \\ \mathcal{M}_1: & \mathbf{X} \sim \mathcal{N}(\mathbf{m},\, \sigma^2\mathbf{I}) \end{array} \qquad \sigma^2 = ?$$

(a) Show that the unknown variance enters into the optimum decision *only* in the threshold term.

(b) In the (happy) situation where the threshold η equals one, show that the optimum test does not depend on σ^2 and that we did not need to know its value in the first place. When will η equal one?

5.38 Unknown Phase Detection
Suppose the received signal consists of a sinusoid in colored Gaussian noise.

$$\mathcal{M}_0 : X_l = A\sin(2\pi f_0 l + \phi) + N_l$$
$$\mathcal{M}_1 : X_l = A\sin(2\pi f_1 l + \phi) + N_l$$

where $l = 0, \ldots, L-1$. The frequencies, which are distinct and equal integer multiples of $1/L$, are known by the receiver but the amplitude and the phase are not. The sinusoids are equally likely to occur.

(a) Derive a receiver for this situation.

(b) What is the optimal choice for the frequencies?

5.39 Phase Déja Vu

There are occasions when the phase is not uniform and we need to test for this occurrence. Assume that the probability density function of the phase has the form

$$p_\Theta(\theta) = \frac{1}{2\pi I_0(a)} e^{a\cos(\theta-\theta_0)}, \quad 0 \le \theta < 2\pi,\ a \ge 0$$

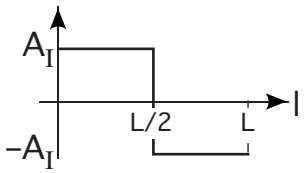
Here, $I_0(a)$ is the modified Bessel function of the first kind, order zero, which serves to normalize the expression so that we have a density. θ_0 is a known phase offset.

(a) Develop a procedure for testing whether a single phase measurement has a uniform distribution or not that does *not* depend on knowing the value of a.

(b) Find an expression for your test's threshold.

5.40 Interference

Wireless communication seldomly occurs without the presence of interference that originates from other communications signals. Suppose a binary communication system uses the antipodal signal set $s_i(l) = (-1)^{i+1}A$, $l = 0,\ldots,L-1$, with $i = 0,1$. An interfering communication system also uses an antipodal signal set having the depicted basic signal. Its amplitude A_I is unknown. The received signal consists of the sum of our signal, the interfering signal, and white Gaussian noise.



(a) What receiver would be used if the bit intervals of the two communication systems were aligned?

(b) How would this receiver change if the bit intervals did *not* align, with the time shift not known? Assume that the receiver knows the time origin of our communications.

5.41 Consider the following composite hypothesis testing problem [89: Prob. 2.5.2].

$$\mathcal{M}_0\colon\ p_X(X) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{X^2}{2\sigma_0^2}\right\}$$

$$\mathcal{M}_1\colon\ p_X(X) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{X^2}{2\sigma_1^2}\right\}$$

where σ_0 is known but σ_1 is known only to be greater then σ_0. Assume that we require that $P_F = 10^{-2}$.

(a) Does an UMP test exist for this problem? If it does, find it.

(b) Construct a generalized likelihood ratio test for this problem. Under what conditions can the requirement on the false-alarm probability be met?

5.42 Assume we have N sensors each determining whether a signal is present in white Gaussian noise or not. The identical signal and noise models apply at each sensor, with the signal having energy E.

(a) What is each sensor's receiver?

(b) Assuming the signal is as likely as not, what is the optimal fusion rule?

(c) Does this distributed detection system yield the same error probabilities as the optimal detector that assimilates all the observations directly?

5.43 Data are often processed "in the field," with the results from several systems sent a central place for final analysis. Consider a detection system wherein each of N field radar systems detects the presence or absence of an airplane. The detection results are collected together so that a final judgment about the airplane's presence can be made. Assume each field system has false-alarm and detection probabilities P_F and P_D respectively.

(a) Find the optimal detection strategy for making a final determination that maximizes the probability of making a correct decision. Assume that the *a priori* probabilities π_0, π_1 of the airplane's absence or presence, respectively, are known.

(b) How does the airplane detection system change when the *a priori* probabilities are not known? Require that the central judgment have a false-alarm probability no bigger than $(P_F)^N$.

5.44 Mathematically, a preconception is a model for the "world" that you believe applies over a broad class of circumstances. Clearly, you should be vigilant and continually judge your assumption's correctness.

Let $\{X_l\}$ denote a sequence of random variables that you believe to be independent and identically distributed with a Gaussian distribution having zero mean and variance σ^2. Elements of this sequence arrive one after the other, and you decide to use the sample average M_l as a test statistic.

$$M_l = \frac{1}{l}\sum_{i=1}^{l} X_i$$

(a) Based on the sample average, develop a procedure that tests for each l whether the preconceived model is correct. This test should be designed so that it continually monitors the validity of the assumptions, and indicates at each l whether the preconception is valid or not. Establish this test so that it yield a constant probability of judging the model incorrect when, in fact, it is actually valid.

(b) To judge the efficacy of this test, assume the elements of the actual sequence have the assumed distribution, but that they are correlated with correlation coefficient ρ. Determine the probability (as a function of l) that your test correctly invalidates the preconception.

(c) Is the test based on the sample average optimal? If so, prove it so; if not, find the optimal one.

5.45 Cars on a Freeway

You have been assigned to develop a simple traffic alert system that warns when the flow of cars differs from a nominal flow. The system measures the separation between cars, which is assumed to have an exponential probability distribution.

$$p_X(x) = \lambda e^{-\lambda x}, \quad x > 0$$

Assuming cars are all the same length, the parameter λ is inversely related to the car's speed v. All cars are assumed to travel at the same speed during any measurement session. Nominal conditions have $\lambda = \lambda_0$. We measure L intervals; if the common speed differs greatly from the nominal, the system is to issue an alert.

(a) Of particular concern to commuters is when a traffic jam is occurring (i.e., $\lambda > \lambda_0$). Develop an alert system for this situation. In other words, find a sufficient statistic and determine how the alert system would use it.

(b) For simplicity, assume $L = 2$. Find an expression for the threshold for an alert system that guarantees that the probability of a mistaken alert is less than 0.25.

(c) Construct a system that issues accurate alerts as soon as possible.

Note: If $p_X(X) = \lambda e^{-\lambda x}\mathrm{u}(x)$, the density of the sum of L statistically independent values is a gamma density. Letting $Y = \sum_{\ell=0}^{L-1} X_\ell$,

$$p_Y(y) = \frac{\lambda^L y^{L-1} e^{-\lambda y}}{(L-1)!}\mathrm{u}(y)$$

5.46 Assume that observations of a sinusoidal signal $s(l) = A\sin(2\pi fl)$, $l = 0,\ldots,L-1$, are contaminated by first-order colored noise as described in the example {182}.

(a) Find the unit-sample response of the whitening filter.

(b) Assuming that the alternative model is the sole presence of the colored Gaussian noise, what is the probability of detection?

(c) How does this probability vary with signal frequency f when the first-order coefficient is positive? Does your result make sense? Why?

5.47 In space-time coding systems, a common bit stream is transmitted over several channels simultaneously but using different signals. $\mathbf{X}^{(k)}$ denotes the signal received from the k^{th} channel, $k = 1,\ldots,K$, and the received signal equals $\mathbf{s}^{(k,i)} + \mathbf{N}^{(k)}$. Here, i equals 0 or 1, corresponding to the bit being transmitted. Each signal has length L. $\mathbf{N}^{(k)}$ denotes a Gaussian random vector with statistically independent components having mean zero and variance σ_k^2 (the variance depends on the channel).

(a) Assuming equally likely bit transmissions, find the minimum probability of error decision rule.

(b) What is the probability that your decision rule makes an error?

(c) Suppose each channel has its *own* decision rule, which is designed to yield the same miss probability as the others. Now what is the minimum probability of error decision rule of the system that combines the individual decisions into one?

5.48 The performance for the optimal detector in white Gaussian noise problems depends only on the distance between the signals. Let's confirm this result experimentally. Define the signal under one hypothesis to be a unit-amplitude sinusoid having one cycle within the 50-sample observation interval. Observations of this signal are contaminated by additive white Gaussian noise having variance equal to 1.5. The hypotheses are equally likely.

(a) Let the second hypothesis be a cosine of the same frequency. Calculate and estimate the detector's false-alarm probability.

(b) Now let the signals correspond to square-waves constructed from the sinusoids used in the previous part. Normalize them so that they have the same energy as the sinusoids. Calculate and estimate the detector's false-alarm probability.

(c) Now let the noise be Laplacian with variance 1.5. Although no analytic expression for the detector performance can be found, do the simulated performances for the sinusoid and the square-wave signals change significantly?

(d) Finally, let the second signal be the negative of the sinusoid. Repeat the calculations and the simulation for Gaussian noise.

5.49 Physical constraints imposed on signals can change what signal set choices result in the best detection performance. Let one of two equally likely discrete-time signals be observed in the presence of white Gaussian noise (variance/sample equals σ^2).

$$\begin{aligned} \mathscr{M}_0 &: X(l) = s^{(0)}(l) + N(l) \\ \mathscr{M}_1 &: X(l) = s^{(1)}(l) + N(l) \end{aligned} \qquad l = 0,\ldots,L-1$$

We are free to choose any signals we like, but there are constraints. Average signal power equals $\sum_l s^2(l)/L$, and peak power equals $\max_l s^2(l)$.

(a) Assuming the average signal power must be less than P_{ave}, what are the optimal signal choices? Is your answer unique?

(b) When the peak power P_{peak} is constrained, what are the optimal signal choices?

(c) If $P_{\text{ave}} = P_{\text{peak}}$, which constraint yields the best detection performance?

5.50 Noisy Signal Templates

In many signal detection problems, the signal itself is not known accurately; for example, the signal could have been the result of a measurement, in which case the "signal" used to specify the observations is actually the actual signal plus noise. We want to determine how the measurement noise affects the detector.

The formal statement of the detection problem is

$$\begin{array}{ll} \mathscr{M}_0: & X(l) = \widetilde{s}^0(l) + N(l) \\ \mathscr{M}_1: & X(l) = \widetilde{s}^1(l) + N(l) \end{array}, \quad l = 0, \ldots, L-1,$$

where $\widetilde{s}^i(l)$ equals $s^i(l) + w(l)$, with $w(l)$ and $N(l)$ comprising white Gaussian noise having variances σ_w^2 and σ_N^2, respectively. We know precisely what the signals $\widetilde{s}^i(l)$ are, but not the underlying "actual" signal.

(a) Find a detector for this problem.

(b) Analyze, as much as possible, how this detector is affected by the measurement noise, contrasting its performance with that obtained when the signals are known precisely.

5.51 One of the more interesting problems in detection theory is determining when the probability distribution of the observations differs from that in other portions of the observation interval. The most common form of the problem is that over the interval $[0, C)$, the observations have one form, and that in the remainder of the observation interval $[C, L-1]$ have a different probability distribution. The change detection problem is to determine whether in fact a change has occurred and, if so, estimate when that change occurs.

To explore the change detection problem, let's explore the simple situation where the mean of white Gaussian noise changes at the C^{th} sample.

$$\mathscr{M}_0: X(l) \sim \mathscr{N}(0, \sigma^2), l = 0, \ldots, L-1$$

$$\mathscr{M}_1: X(l) \sim \begin{cases} \mathscr{N}(0, \sigma^2), & l = 0, \ldots, C-1 \\ \mathscr{N}(m, \sigma^2), & l = C, \ldots, L-1 \end{cases}$$

The observations in each case are statistically independent of the others.

(a) Find a detector for this change problem when m is a known positive number.

(b) Find an expression for the threshold in the detector using the Neyman-Pearson criterion.

(c) How does the detector change when the value of m is not known?

5.52 Noise generated by a system having zeros complicates the calculations for the colored noise detection problem. To illustrate these difficulties, assume the observation noise is produced as the output of a filter governed by the difference equation

$$N(l) = aN(l-1) + w(l) + bw(l-1), \quad a \neq -b,$$

where $w(l)$ is white, Gaussian noise. Assume an observation duration sufficiently long to capture the colored effects.

(a) Find the covariance matrix of this noise process.

(b) Calculate the Cholesky factorization of the covariance matrix.

(c) Find the unit-sample response of the optimal detector's whitening filter. If it weren't for the finite observation interval, would it indeed have an infinite-duration unit-sample response as claimed on Page 183? Describe the edge-effects of your filter, contrasting them with the case when $b = 0$.

5.53 It is frequently claimed that the relation between noise bandwidth and reciprocal duration of the observation interval play a key role in determining whether DFT values are approximately uncorrelated. While the statements sound plausible, their veracity should be checked. Let the covariance function of the observation noise be $K_N(l) = a^{|l|}$.

(a) How is the bandwidth (defined by the half-power point) of this noise's power spectrum related to the parameter a? How is the duration (defined to be two time constants) of the covariance function related to a?

(b) Find the variance of the length-L DFT of this noise process as a function of the frequency index k. This result should be compared with the power spectrum calculated in part (a); they *should* resemble each other when the "memory" of the noise—the duration of the covariance function—is much less than L while demonstrating differences as the memory becomes comparable to or exceeds L.

(c) Calculate the covariance between adjacent frequency indices. Under what conditions will they be approximately uncorrelated? Relate your answer to the relations of a to L found in the previous part.

5.54 The results derived in Problem 5.53 assumed that a length-L Fourier Transform was computed from a length-L segment of the noise process. What will happen if the transform has length $2L$ with the observation interval remaining unchanged?

(a) Find the variance of DFT values at index k.

(b) Assuming the conditions in Problem 5.53 for uncorrelated adjacent samples, now what is the correlation between adjacent DFT values?

5.55 In a discrete-time detection problem, one of two, equally likely, length-L sinusoids (each having a frequency equal to a known multiple of $1/L$) is observed in additive colored Gaussian noise. Signal amplitudes are also unknown by the receiver. In addition, the power spectrum of the noise is uncertain: what is known is that the noise power spectrum is broadband, and varies gently across frequency. Find a detector that performs well for this problem. What notable properties (if any) does your receiver have?

5.56 A sampled signal is suspected of consisting of a periodic component and additive Gaussian noise. The signal, if present, has a known period L. The number of samples equals N, a multiple of L. The noise is white and has known variance σ^2. A consultant (you!) has been asked to determine the signal's presence.

(a) Assuming the signal is a sinusoid with unknown phase and amplitude, what should be done to determine the presence of the sinusoid so that a false-alarm probability criterion of 0.1 is met?

(b) Other than its periodic nature, now assume that the signal's waveform is unknown. What computations must the optimum detector perform?

5.57 The QAM (Quadrature Amplitude Modulation) signal set consists of signals of the form

$$s_i(l) = A_i^c \cos(2\pi f_c l) + A_i^s \sin(2\pi f_c l) \ ,$$

where A_i^c and A_i^s are amplitudes that define each element of the signal set. These are chosen according to design constraints. Assume the signals are observed in additive Gaussian noise.

(a) What is the optimal amplitude choice for the binary and quaternary (four-signal) signal sets when the noise is white and the signal energy is constrained ($\sum_l s_i^2(l) < E$)? Comment on the uniqueness of your answers.

(b) Describe the optimal binary QAM signal set when the noise is colored.

(c) Now suppose the *peak* amplitude ($\max_l |s_i(l)| < A_{\max}$) is constrained. What are the optimal signal sets (both binary and quaternary) for the white noise case? Again, comment on uniqueness.

5.58 I confess that unknown-delay problem given on Page 188 is not terribly relevant to active sonar and radar ranging problems. In these cases, the signal's delay, measured with respect to its emission by a source, presents the round-trip propagation time from the source to the energy-reflecting object and back. Hence, delay is proportional to twice the range. What makes the example overly simplistic is the independence of signal amplitude on delay.

(a) Because of attenuation due to spherical propagation, show that the received signal energy is inversely related to the fourth power of the range. This result is known as the *radar equation*.

(b) Derive the detector that takes the dependence of delay and amplitude into account, thereby optimally determining the presence of a signal in active radar/sonar applications, and produces a delay estimate, thereby simultaneously providing the object's range. Not only determine the detector's structure, but also how the threshold and false-alarm probability are related.

(c) Does the object's reflectivity need to be known to implement your detector?

5.59 Checking for Repetitions

Consider the following detection problem.

$$\begin{aligned} \mathscr{M}_0: & \quad \mathbf{X}_1 = \mathbf{s} + \mathbf{N}_1 \\ & \quad \mathbf{X}_2 = \mathbf{s} + \mathbf{N}_2 \\ \mathscr{M}_1: & \quad \mathbf{X}_1 = \mathbf{s}_1 + \mathbf{N}_1 \\ & \quad \mathbf{X}_2 = \mathbf{s}_2 + \mathbf{N}_2 \end{aligned}$$

Here, the two observations either contain the same signal or they contain different ones. The noise vectors N_1 and N_2 are statistically independent of each other and identically distributed, with each being Gaussian with zero mean and covariance matrix $\mathbf{K} = \sigma^2\mathbf{I}$.

(a) Find the decision rule that minimizes the false-alarm probability when the miss probability is required to be less than $1-\beta$.

(b) Now suppose *none* of the signals is known. All is that is known is that under $\mathscr{M}_0$, the signals are the same and that under $\mathscr{M}_1$ they are different. What is the optimal decision rule under these conditions?

5.60 *CFAR* detectors are extremely important in applications because they automatically adapt to the value of noise variance during the observations, allowing them to be used in varying noise situations. However, as described in Problem 5.58, unknown delays must be also considered in realistic problems.

(a) Derive a *CFAR* detector that also takes unknown signal delay into account.

(b) Show that your detector *automatically* incorporates amplitude uncertainties.

(c) Under the no-signal model, what is the distribution of the sufficient statistic?

5.61 A sampled signal is suspected of consisting of a periodic component and additive Gaussian noise. The signal, if present, has a known period N. The number of samples equals L, a multiple of N. The noise is white and has known variance σ^2. A consultant (you!) has been asked to determine the signal's presence.

(a) Assuming the signal is a sinusoid with unknown phase and amplitude, what should be done to determine the presence of the sinusoid so that a false-alarm probability criterion of 0.1 is met?

(b) Other than its periodic nature, now assume that the signal's waveform is unknown. What computations must the optimum detector perform?

5.62 Delegating Responsibility

Modern management styles tend to want decisions to be made locally (by people at the scene) rather than by "the boss." While this approach might be considered more democratic, we should understand how to make decisions under such organizational constraints and what the performance might be.

Let three "local" systems separately make observations. Each local system's observations are identically distributed and statistically independent of the others, and based on the observations, each system decides which of two models applies best. The judgments are relayed to the central manager who must make the final decision. Assume the local observations consist either of white Gaussian noise or of a signal having energy E to which the same white Gaussian noise has been added. The signal energy is the same at each local system. Each local decision system must meet a performance standard on the probability it declares the presence of a signal when none is present.

(a) What decision rule should each local system use?

(b) Assuming the observation models are equally likely, how should the central management make its decision so as to minimize the probability of error?

(c) Is this decentralized decision system optimal (*i.e.*, the probability of error for the final decision is minimized)? If so, demonstrate optimality; if not, find the optimal system.

5.63 The additive noise in a detection problem consists of a sequence of statistically independent Laplacian random variables. The probability density of $N(l)$ is therefore

$$p_{N(l)}(N) = \frac{1}{2}e^{-|N|}.$$

The two possible signals are constant throughout the observation interval, equaling either $+1$ or -1.

(a) Find the optimum decision rule which could be used on a *single* value of the observation signal.

(b) Find an expression for the threshold in your rule as a function of the false-alarm probability.

(c) What is the threshold value for the "symmetric" error situation ($P_F = 1 - P_D$)?

(d) Now assume that two values of the observations are used in the detector (patience; to solve this problem, we need to approach it gradually). What is the decision rule for symmetric errors? How does this result generalize to L observations?

5.64 The threshold for non-parametric model evaluation is computed (somewhat glibly) on Page 201 using the Central Limit Theorem. As warned in Chap. 2, such applications of the Central Limit Theorem need to employed carefully.

(a) The critical parameters—$\mathsf{E}[\mathrm{u}(x)]$ and $\mathsf{var}[\mathrm{u}(x)]$—of the Central Limit Theorem approximation are claimed to equal $\frac{1}{2}$ and $\frac{1}{4}$ respectively under model $\mathscr{M}_0$. Verify this claim.

(b) How is the threshold in the non-parametric hypothesis test related to the number of observations from the point of view of the Central Limit Theorem?

(c) Contrast this answer with that determined by the non-parametric test. Which constraint dominates under what conditions?

(d) How many observations are needed in a non-parametric test to achieve a false-alarm rate of 10^{-2}? of 10^{-6}?

5.65 What are the ranges of signal waveforms and probability density functions consistent with the appropriate model for uncertainty about a nominal?

(a) When considering waveform uncertainties (Page 202), the uncertainty was modeled as an additive corrupting signal that has bounded energy. Assuming a sinusoidal nominal consisting of eight samples per period, sketch the worst-case and best-case signals for the detection problem described in the text.

(b) What other signals are consistent with this model? Do they fall within the envelopes of the best-case and worst-case signals? If so prove it; if not, find a counter example and provide some method of finding the others.

(c) The ε-contamination model for uncertain densities was somewhat different because we needed to maintain unit area (Page 203). Characterize the range of densities consistent with this model.

5.66 In the partially known signal waveform problem discussed on Page 202, we assumed that there was no uncertainty in the "no-signal" model $\mathscr{M}_0$. A residual signal corrupting the observations may well be present no matter what signal model is actually true. For example, hum (power-line noise) could be present no matter what signal is received.

(a) Re-derive the robust detector for the situation where a zero-energy signal and a non-zero energy signal can both be corrupted by a signal having energy E^c.

(b) Under what conditions will a solution exist?

(c) Contrast your detector with that found in the text. What are the worst-case signals? What signal is used in the matched-filter in the white noise case? Does the usual matched filter remain robust?

5.67 The detector robust under signal waveform uncertainties was only briefly described in the text (see Page 202). This result has several interesting properties that warrant further exploration.

(a) Explicitly derive the detector using the Lagrange multiplier technique outlined in the text. Find the value of the Lagrange multiplier for the white-noise case.

(b) Don't take the author's word that the worst-case signal is indeed that derived. By computing the probability of detection for a matched-filter detector that uses $s^o(l)$, find the worst-case signal $s^w(l)$—the one that minimizes P_D—that satisfies the constraint $s^w(l) = s^o(l) + c(l)$, $\sum c^2(l) \leq E^c$. Do *not* assume the noise is white.

(c) The solution shown on Page 202 differs from the colored-noise detector that ignores signal uncertainties. By contrasting the signals to which the observations are matched, describe the differences. When are they similar? very different?

5.68 In robust detection problems in which we have *signal uncertainties*, the so-called "minimax" approach is to find the worst-case signals among received signal possibilities (the pair that would yield the worst performance if they were indeed transmitted), then use the likelihood ratio based on these signals instead of what is actually transmitted. Assume the signal transmissions are observed in the presence of additive white Gaussian noise. The received signal $\tilde{\mathbf{s}}_i$ deviates from the nominal one $\mathbf{s}_i$ by a finite mean-squared error: $\| \tilde{\mathbf{s}}_i - \mathbf{s}_i \|^2 \leq \varepsilon^2$.

Prove or disprove: The receiver obtained when there is no uncertainty is robust.

5.69 Consider a two-model problem where signals are observed in the presence of additive *Laplacian* noise having samples statistically independent of each other and the signal.

$$\begin{array}{ll} \mathscr{M}_0: & X(l) = s_0(l) + N(l) \\ \mathscr{M}_1: & X(l) = s_1(l) + N(l) \end{array} , \; l = 0, \ldots, L-1$$

To design an optimal signal set for this problem, we constrain each signal's energy to be less than E.

(a) What signal choices provide optimal performance (the smallest error probabilities) when the signal-to-noise ratio is small?

(b) When the signal-to-noise ratio is large, do the choices change? If so, determine what choices, if any, would be optimal regardless of the signal-to-noise ratio.

5.70 In Poisson channels, such as photon-limited optical communication systems, the received signal consists of the number of photons and when they occurred. The joint distribution of these quantities has the form (2.6) {16}

$$p_{N_{0,T},\mathbf{W}}(n;\mathbf{w}) = \left(\prod_{k=1}^{n} \lambda(w_k) \right) \exp\left\{ - \int_0^T \lambda(\alpha)\, d\alpha \right\} ,$$

where $\lambda(t)$ denotes the Poisson process's time-varying intensity, which *must* be non-negative. In optical digital communication, the signal sets consist of designed intensities, and these are typically received in the presence of a background light level. Phrased in terms of a model, $\mathscr{M}_i$: $\lambda(t) = \lambda_i(t) + \lambda_d$. In designing signal sets for the optical channel, the intensity's integral is typically bounded: $\int_0^T \lambda_i(\alpha)\, d\alpha < \Lambda_{\max}$. What is the optimal binary signal set $\{\lambda_0(t), \lambda_1(t)\}$ for the optical channel?

Appendix A

Probability Distributions

The table on the following pages details common probability distributions.

Discrete Random Variables

Name	*Probability*	*Mean*	*Variance*	*Relationships*
Discrete Uniform	$\begin{cases} \frac{1}{N-M+1} & M \le n \le N \\ 0 & \text{otherwise} \end{cases}$	$\frac{M+N}{2}$	$\frac{(N-M+2)(N-M)}{12}$	
Bernoulli or Binary	$\Pr(n=0) = 1-p$ $\Pr(n=1) = p$	p	$p(1-p)$	
Geometric	$(1-p)p^n$, $n \ge 0$	$p/1-p$	$p/(1-p)^2$	
Binomial	$\binom{N}{n} p^n (1-p)^{N-n}$, $n = 0, \ldots, N$	Np	$Np(1-p)$	Sum of N IID Bernoulli
Negative Binomial	$\binom{n-1}{N-1} p^N (1-p)^{n-N}$, $n \ge N$	N/p	$N(1-p)/p^2$	
Poisson	$\frac{\lambda^n e^{-\lambda}}{n!}$, $n \ge 0$	λ	λ	
Hypergeometric	$\frac{\binom{a}{n}\binom{b}{N-n}}{\binom{a+b}{N}}$, $n = 0, \ldots, N$; $0 \le N \le a+b$	$\frac{Na}{a+b}$	$\frac{Nab(a+b-N)}{(a+b)^2(a+b-1)}$	
Logarithmic	$\frac{-p^n}{n\log(1-p)}$, $n > 0$	$\frac{-p}{(1-p)\log(1-p)}$	$\frac{-p[p+\log(1-p)]}{(1-p)\log q}$	

Continuous Random Variables

Gaussian (Normal)	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$	m	σ^2	
Bivariate Gaussian	$\frac{1}{2\pi(1-\rho^2)^{1/2}\sigma_x\sigma_y} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-m_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right) + \left(\frac{y-m_y}{\sigma_y}\right)^2\right]\right\}$			
		$\mathsf{E}[x] = m_x$, $\mathsf{E}[y] = m_y$	$\mathsf{var}[x] = \sigma_x^2$, $\mathsf{var}[y] = \sigma_y^2$, $\mathsf{E}[xy] = m_x m_y + \rho\sigma_x\rho\sigma_y$	ρ: correlation coefficient

Continued on next page

Name	*Density*	*Mean*	*Variance*	*Relationships*
Conditional Gaussian	$p(x\|y)=\frac{1}{\sqrt{2\pi(1-\rho)^2\sigma_x^2}}\exp\left\{-\frac{\left(x-m_x-\frac{\rho\sigma_x}{\sigma_y}(y-m_y)\right)^2}{2\sigma_x^2(1-\rho^2)}\right\}$			
		$m_x+\frac{\rho\sigma_x}{\sigma_y}(y-m_y)$	$\sigma_x^2(1-\rho^2)$	
Multivariate Gaussian	$\frac{1}{(\det(2\pi\mathbf{K}))^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\right\}$	$\mathbf{m}$	$\mathbf{K}$	
Generalized Gaussian	$\frac{1}{2\Gamma(1+1/r)A(r)}\exp\left\{-\left\|\frac{x-m}{A(r)}\right\|^r\right\}$	m	σ^2	$A(r)=\left[\frac{\sigma^2\Gamma(1/r)}{\Gamma(3/r)}\right]^{1/2}$
Chi-Squared (χ_ν^2)	$\frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\frac{\nu}{2}-1}e^{-x/2},\ 0\le x$	ν	2ν	$\chi_\nu^2=\sum_{i=1}^{\nu}X_i^2$, X_i IID $\mathcal{N}(0,1)$
Noncentral Chi-Squared ($\chi_\nu'^2(\lambda)$)	$\frac{1}{2}(x/\lambda)^{(\nu-2)/4}I_{(\nu-2)/2}(\sqrt{\lambda}x)e^{-1/2(\lambda+x)}$	$\nu+\lambda$	$2(\nu+2\lambda)$	$\chi_\nu'^2=\sum_{i=1}^{\nu}X_i^2$, X_i IID $\mathcal{N}(m_i,1)$ $\lambda=\sum_{i=1}^{\nu}m_i^2$
Chi χ_ν	$\frac{x^{\nu-1}e^{-x^2/2}}{2^{\nu/2-1}\Gamma\left(\frac{\nu}{2}\right)},\ 0\le x$	$\sqrt{2}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\nu\right)}$	$\frac{2\left[\Gamma\left(\frac{1}{2}\nu\right)\Gamma\left(\frac{1}{2}\nu+1\right)-\Gamma^2\left(\frac{\nu+1}{2}\right)\right]}{\Gamma^2\left(\frac{1}{2}\nu\right)}$	$\chi_\nu=\sqrt{\chi_\nu^2}$
Student's t	$\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	0	$\frac{\nu}{\nu-2},\ 2<\nu$	
Beta $\beta_{m,n}$	$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}x^{m/2-1}(1-x)^{n/2-1},\ 0<x<1,$ $0<a,b$	$\frac{m}{m+n}$	$\frac{2mn}{(m+n)^2(m+n+2)}$	$\beta_{m,n}=\frac{\chi_m^2}{\chi_m^2+\chi_n^2}$
F Distribution	$\frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2}\frac{x^{(m-2)/2}}{[1+(m/n)x]^{(m+n)/2}},\ 0\le x;1\le m,n$			
		$\frac{n}{n-2},\ n>2$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)},\ n>4$	$F_{m,n}=\frac{\chi_m^2/m}{\chi_n^2/n}$

Continued on next page

Name	*Density*	*Mean*	*Variance*	*Relationships*
Non-central F $F'_{m,n}(\lambda)$	$\sum_{k=0}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^k}{k!} e^{-\frac{\lambda}{2}} p_{\beta_{\frac{m}{2}+k,\frac{n}{2}}}\left(\frac{mx}{mx+n}\right)$	$\frac{n}{n-2}, n>2$	$2\left(\frac{n}{m}\right)^2 \frac{(m+\lambda)^2+(m+2\lambda)(n-2)}{(n-2)^2(n-4)}$, $n>4$	$F'_{m,n}(\lambda) = \frac{{\chi'_m}^2(\lambda)/m}{\chi^2_n/n}$
Wishart $\mathbf{W}_M(N,\mathbf{K})$	$\frac{(\det[\mathbf{w}])^{\frac{N-M-1}{2}}}{2^{\frac{NM}{2}}\Gamma_M(\frac{N}{2})(\det[\mathbf{K}])^{\frac{K}{2}}} e^{-\mathrm{tr}[\mathbf{K}^{-1}\mathbf{w}]/2}$, with $\Gamma_M(\frac{N}{2}) = \pi^{M(M-1)/4} \times \prod_{m=0}^{M-1} \Gamma\left(\frac{N}{2}-\frac{m}{2}\right)$	$N\mathbf{K}$	$\mathrm{cov}[\mathbf{W}_{ij}\mathbf{W}_{kl}] = N\cdot(\mathbf{K}_{ik}\mathbf{K}_{jl}+\mathbf{K}_{il}\mathbf{K}_{jk})$	$\mathbf{W}_M(N,\mathbf{K}) = \sum_{n=1}^{N} \mathbf{X}_n\mathbf{X}'_n$, $\mathbf{X}_n \sim \mathcal{N}(\mathbf{0},\mathbf{K})$ $\dim[\mathbf{X}] = M$
Uniform	$\frac{1}{b-a}, a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Triangular	$2x/a \quad 0 \le x \le a$ $2(1-x)/(1-a) \quad a \le x \le 1$	$\frac{1+a}{3}$	$\frac{1-a+a^2}{18}$	
Exponential	$\lambda e^{-\lambda x}, 0 \le x$	$1/\lambda$	$1/\lambda^2$	
Lognormal	$\frac{1}{\sqrt{2\pi\sigma^2 y^2}} \exp\left\{-\frac{1}{2}\left(\frac{\log y - m}{\sigma}\right)^2\right\}, y>0$	$e^{m+\frac{\sigma^2}{2}}$	$e^{2m}\left(e^{2\sigma^2}-e^{\sigma^2}\right)$	
Rayleigh	$2aye^{-ay^2}, y \ge 0$	$\sqrt{\frac{\pi}{4a}}$	$\frac{1}{a}\left(1-\frac{\pi}{4}\right)$	$Y = \sqrt{X_1^2+X_2^2}$, X_i IID $\mathcal{N}(0,\sigma^2)$, $a = \frac{1}{2\sigma^2}$
Maxwell	$\sqrt{\frac{2}{\pi}} a^{-3/2} y^2 e^{-y^2/2a}, 0 < y$	$\sqrt{\frac{8}{\pi a}}$	$\left(3-\frac{8}{\pi}\right)a^{-1}$	$Y = \sqrt{X_1^2+X_2^2+X_3^2}$, X_i IID $\mathcal{N}(0,\sigma^2)$, $a = \frac{1}{\sigma^2}$

Continued on next page

Name	*Density*	*Mean*	*Variance*	*Relationships*
Laplacian	$\frac{1}{\sqrt{2\sigma^2}}\exp\left\{-\frac{\lvert x-m\rvert}{\sqrt{\sigma^2/2}}\right\}$	m	σ^2	
Gamma	$\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$, $x>0$, $a,b>0$	$\frac{a}{b}$	$\frac{a}{b^2}$	
Weibull	$abx^{b-1}e^{-ax^b}$, $x>0$, $a,b>0$	$(1/a)^{1/b}\cdot\Gamma(1+1/b)$	$a^{-2/b}\cdot\left[\Gamma(1+2/b)-\Gamma^2(1+1/b)\right]$	
Arc-Sine	$\frac{1}{\pi\sqrt{x(1-x)}}$, $0<x<1$	$\frac{1}{2}$	$\frac{1}{8}$	
Sine Amplitude	$\frac{1}{\pi\sqrt{1-x^2}}$, $\lvert x\rvert<1$	0	$\frac{1}{2}$	
Circular Normal	$\frac{e^{a\cos(x-m)}}{2\pi I_0(a)}$, $-\pi<x\leq\pi$	m		
Cauchy	$\frac{a/\pi}{(x-m)^2+a^2}$	m (from symmetry arguments)	∞	
Logistic	$\frac{e^{-(x-m)/a}}{a\left[1+e^{-(x-m)/a}\right]^2}$, $a>0$	m	$\frac{a^2\pi^2}{3}$	
Gumbel	$\frac{e^{-(x-m)/a}}{a}\exp\left\{-e^{-(x-m)/a}\right\}$, $a>0$	$m+a\gamma$	$\frac{a^2\pi^2}{6}$	
Pareto	$\frac{ab^a}{x^{1-a}}$, $a>0$; $x\geq b>0$	$\frac{ab}{a-1}$, $a>1$	$\frac{ab^2}{(a-2)(a-1)^2}$, $a>2$	

Appendix B

Matrix Theory

B.1 Basic Definitions

An $m \times n$ *matrix* $\mathbf{A}$ is a rectangular (square if $n = m$) arrangement of scalar elements A_{ij} (i^{th} row, j^{th} column).

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}$$

When the matrix is square, the *main diagonal* extends from the upper left corner to the lower right and consists of the elements A_{ii}. The *cross diagonal* is the opposite diagonal and consists of the elements $A_{n-i+1,i}$. A *vector* is common terminology for a column ($m \times 1$) matrix. The *dimension* of the vector equals m. Rectangular matrices are (usually) denoted by boldface uppercase letters ($\mathbf{A}, \mathbf{B}, \ldots$) and vectors by boldface lowercase letters ($\mathbf{a}, \mathbf{b}, \ldots$). The dimension of a vector is expressed as $\dim(\mathbf{a})$.

To add matrices, the number of rows in each matrix as well as the number of columns must agree. The sum of $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ is defined to be an $m \times n$ matrix $\mathbf{C}$ whose elements are

$$C_{ij} = A_{ij} + B_{ij}.$$

Matrix addition is commutative ($\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$). The product $\mathbf{AB}$ of two matrices $\mathbf{A}$ and $\mathbf{B}$ is only defined if the number of columns of $\mathbf{A}$ equals the number of rows of $\mathbf{B}$. Thus, if $\mathbf{A}$ represents an $l \times m$ matrix and $\mathbf{B}$ an $m \times n$ matrix, the result is an $l \times n$ matrix $\mathbf{C}$, each term of which is defined to be

$$C_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}, \quad i = 1, \ldots, l; j = 1, \ldots, n.$$

Clearly, the matrix product is not necessarily commutative ($\mathbf{AB} \neq \mathbf{BA}$), but is distributive [$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$] and associative [$\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$].

Several methods of rearranging the elements of a matrix are frequently used. The *complex conjugate* of the matrix $\mathbf{A}$ is denoted by $\mathbf{A}^*$ and consists of the complex conjugates of the elements of $\mathbf{A}$.

$$[\mathbf{A}^*]_{ij} = A_{ij}^*$$

The *transpose* of an $m \times n$ matrix $\mathbf{A}$ is an $n \times m$ matrix $\mathbf{A}^t$ whose rows are the columns of $\mathbf{A}$.

$$\left[\mathbf{A}^t\right]_{ij} = A_{ji}$$

The transpose of a product of two matrices equals the product of their transposes, but in reversed order.

$$(\mathbf{AB})^t = \mathbf{B}^t \mathbf{A}^t$$

This property applies whether the matrices are square or not. The *conjugate transpose* (sometimes known as the Hermitian transpose) of $\mathbf{A}$ is denoted by $\mathbf{A}'$ and equals $(\mathbf{A}^*)^t$.

$$\left[\mathbf{A}'\right]_{ij} = A_{ji}^*$$

B.2 Basic Matrix Forms

The relations between the values of the elements of a matrix define special matrix cases. Matrices having special internal structure are important to recognize in manipulating matrix expressions.

- A *diagonal* matrix, denoted by $\text{diag}[A_{11}, A_{22}, \ldots, A_{nn}]$, has non-zero entries only along the main diagonal ($i = j$) of the matrix.
$$\begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & 0 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & A_{nn} \end{bmatrix}$$
- A *block diagonal* matrix has square matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N$ on the diagonal, with zero-valued entries elsewhere. Note that these matrices do not necessary share a common size.
$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \cdots \\ \vdots & & \ddots & \vdots \\ \cdots & \cdots & \mathbf{0} & \mathbf{A}_N \end{bmatrix}$$
- The *identity* matrix, denoted by $\mathbf{I}$, is a special diagonal matrix having unity on the main diagonal.
$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$
The identity matrix is so-named because it is the multiplicative identity for square matrices ($\mathbf{IA} = \mathbf{AI} = \mathbf{A}$).
- The ordering of the rows or columns of a square matrix can be reversed by pre- or post-multiplication by the *exchange* matrix $\mathbf{J}$.
$$\mathbf{J} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & 1 & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}$$
- A *lower triangular* matrix $\mathbf{L}$ has non-zero values on and "below" its main diagonal.
$$\mathbf{L} = \begin{bmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & 0 & \vdots \\ \vdots & \vdots & \ddots & \\ L_{n1} & L_{n2} & \cdots & L_{nn} \end{bmatrix}$$
An upper triangular matrix can be similarly defined.
- A *Vandermonde* matrix consists of columns containing geometric sequences.
$$\mathbf{A} = \begin{bmatrix} 1 & \cdots & 1 \\ a_0 & \cdots & a_{n-1} \\ a_0^2 & \cdots & a_{n-1}^2 \\ \vdots & & \vdots \\ a_0^{m-1} & \cdots & a_{n-1}^{m-1} \end{bmatrix}$$

One special case of a square Vandermonde matrix is the DFT matrix $\mathbf{W}$ in which the elements are $a_k = \exp\{-j2\pi k/m\}$. The discrete Fourier transform of a vector $\mathbf{x}$ can be expressed as

$$\text{DFT}[\mathbf{x}] = \mathbf{W}\mathbf{x}.$$

- A *symmetric* matrix equals its transpose ($\mathbf{A}^t = \mathbf{A}$). A *conjugate symmetric* (Hermitian) matrix equals its own conjugate transpose ($\mathbf{A}' = \mathbf{A}$). Correlation matrices, which have the form $\mathbf{A} = \mathsf{E}[\mathbf{x}\mathbf{x}']$, are Hermitian.
- A *Toeplitz* matrix has constant values along each of its diagonals ($A_{ij} = a_{i-j}$).

$$\begin{bmatrix} a_0 & a_1 & \cdots & a_{n-1} \\ a_{-1} & a_0 & a_1 & \\ \vdots & \ddots & \ddots & \\ a_{-(n-1)} & a_{-(n-2)} & \cdots & a_0 \end{bmatrix}$$

  In time series, correlation matrices are not only Hermitian, they are also Toeplitz. Hermitian Toeplitz matrices have no more than n unique elements in the matrix. Because of this extensive redundancy of values, many efficient techniques for computing expressions involving them have been developed.
- Constant values along the cross-diagonals of a square matrix define a *Hankel* matrix.

$$A_{ij} = a_{i+j-2} \qquad \mathbf{A} = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{n-1} \\ a_1 & a_2 & \cdots & \iddots & a_n \\ a_2 & \iddots & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & & a_{2n-3} \\ a_{n-1} & a_n & \cdots & a_{2n-3} & a_{2n-2} \end{bmatrix}$$

  All Hankel matrices are symmetric. If $\mathbf{A}$ is a Hankel matrix, then both $\mathbf{JA}$ and $\mathbf{AJ}$ are Toeplitz matrices. Hankel matrices are frequently used to express in matrix form the convolution of a sequence with a finite-duration unit sample response. For example, if $y(n) = \sum_{m=n-(M-1)}^{n} h(n-m)x(m)$ for $n = M-1,\ldots,N-1$, the matrix form would be

$$\begin{bmatrix} x(0) & x(1) & \cdots & x(M-1) \\ x(1) & x(2) & \cdots & x(M) \\ \vdots & \vdots & & \vdots \\ x(N-M) & x(N-M+1) & \cdots & x(N-1) \end{bmatrix} \begin{bmatrix} h(M-1) \\ h(M-2) \\ \vdots \\ h(0) \end{bmatrix} = \begin{bmatrix} y(M-1) \\ y(M) \\ \vdots \\ y(N-1) \end{bmatrix}.$$

- A square matrix is said to be *circulant* if each row is a circular rotation of the previous one.

$$\begin{bmatrix} a_0 & a_1 & \cdots & a_{n-1} \\ a_{n-1} & a_0 & a_1 & \cdots \\ a_{n-2} & a_{n-1} & a_0 & \cdots \\ \vdots & \vdots & & \\ a_1 & a_2 & \cdots & a_0 \end{bmatrix}$$

  All circulant matrices are Toeplitz; circulant matrices are symmetric only when the sequence $a_1, a_2, \ldots$ is even: $a_i = a_{n-i}, i = 1,\ldots,n/2$ (a_0 is arbitrary).
- An *orthogonal* matrix $\mathbf{A}$ satisfies $\mathbf{A}^t\mathbf{A} = \mathbf{I}$: its transpose corresponds to its inverse.
- $\mathbf{A}$ is *unitary* if it satisfies $\mathbf{A}'\mathbf{A} = \mathbf{I}$: its conjugate-transpose equals its inverse.
- The square matrix $\mathbf{A}$ is said to be a *projection* matrix if $\mathbf{A}^2 = \mathbf{A}$. Because of this property, $\mathbf{A}^m = \mathbf{A}^n$ for all choices of positive integers n, m. The reason for this name will be given once eigenvalues are defined (§B.6 {245}).

- A *permutation* matrix is a square matrix that has exactly one non-zero value equal to one in each row and each column. The result of applying a permutation matrix to a vector is to permute (shuffle) the vector's values. For example, the following permutation matrix when applied to a vector results in

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{Px} = \begin{bmatrix} x_2 \\ x_4 \\ x_1 \\ x_3 \end{bmatrix}$$

  Note that a projection matrix is a special case of an orthogonal matrix: $\mathbf{P}^t\mathbf{P} = \mathbf{I}$.
- A matrix is said to have a *null space* if a set of non-zero vectors $\mathbf{x}$ exist that satisfy $\mathbf{Ax} = \mathbf{0}$. These vectors would thereby be orthogonal to the rows of $\mathbf{A}$. The matrix could be square or, more interestingly, it could be rectangular, having more columns than rows. The rows would then define a collection of vectors that represent all vectors for which $\mathbf{Ax} \neq 0$.
- A square matrix is said to be *positive-definite* if for every vector $\mathbf{x} \neq \mathbf{0}$, the quadratic form (see §B.4) $\mathbf{x}'\mathbf{Ax}$ is a strictly positive scalar: $\mathbf{x}'\mathbf{Ax} > 0$. A square matrix is said to be *non-negative definite* if for every vector $\mathbf{x} \neq \mathbf{0}$, the quadratic form $\mathbf{x}'\mathbf{Ax} \geq 0$.
- The vector $\mathbf{y}$ is said to be in the *range* of the matrix $\mathbf{A}$ if it can be expressed as $\mathbf{Ax} = \mathbf{y}$ for some $\mathbf{x}$.

B.3 Operations on Matrices

Several operations on matrices and between them are so commonly used that specific terminology has evolved.

- The *inner product* $\mathbf{x}'\mathbf{y}$ between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined to be the scalar expressed by*

$$\mathbf{x}'\mathbf{y} = \sum_{i=1}^{n} x_i^* y_i.$$

  Since an inner product is a scalar, expressions involving inner products are frequently rearranged by noting that the transpose of a scalar is the same scalar. Therefore, $(\mathbf{x}'\mathbf{y})^t = \mathbf{y}^t\mathbf{x}^* = \mathbf{x}'\mathbf{y}$. Two vectors are said to be *orthogonal* if their inner product is zero.
- The *outer product* $\mathbf{xy}'$ between two vectors x (dimension m) and y (dimension n) is an $m \times n$ matrix whose elements are

$$\left[\mathbf{xy}'\right]_{ij} = x_i y_j^*.$$

- The *Kronecker product* $\mathbf{A} \otimes \mathbf{B}$ between two matrices $\mathbf{A}$ ($m_\mathbf{A} \times n_\mathbf{A}$) and $\mathbf{B}$ ($m_\mathbf{B} \times n_\mathbf{B}$) is the $m_\mathbf{A} m_\mathbf{B} \times n_\mathbf{A} n_\mathbf{B}$ matrix given by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} A_{11}\mathbf{B} & A_{12}\mathbf{B} & \cdots & A_{1n_\mathbf{A}}\mathbf{B} \\ A_{21}\mathbf{B} & A_{22}\mathbf{B} & & A_{2n_\mathbf{A}}\mathbf{B} \\ \vdots & & \ddots & \vdots \\ A_{m_\mathbf{A}1}\mathbf{B} & A_{m_\mathbf{A}2}\mathbf{B} & \cdots & A_{m_\mathbf{A}n_\mathbf{A}}\mathbf{B} \end{bmatrix}.$$

  The matrix $\mathbf{B}$ multiplied by the scalars A_{ij} is thus repeated throughout the matrix. The Kronecker product of two positive-definite matrices is also positive definite.
- The *inverse* of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{-1}$ which satisfies $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. The inverse of a matrix is not guaranteed to exist. Numerous conditions on the inverse's existence are detailed in the following. When it does exist, the following properties hold.

* In more abstract settings, the inner product between two vectors x and y is denoted by $\langle x, y \rangle$. There, the inner product becomes any mapping of two vectors to a scalar that satisfies three properties.

(i) $\langle y, x \rangle = \langle x, y \rangle^*$
(ii) $\langle ax + by, z \rangle = a^* \langle x, z \rangle + b^* \langle y, z \rangle$, a, b scalars
(iii) $\langle x, x \rangle > 0$, $x \neq 0$

1. If $\mathbf{A}$, $\mathbf{B}$ are invertible matrices, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
2. If $\mathbf{A}$, $\mathbf{B}$ are invertible matrices, $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$
3. Assuming that all inverses exist where used, the inverse of a sum of matrices can be written several useful ways.

$$\begin{aligned}(\mathbf{A}+\mathbf{BCD})^{-1} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{DA}^{-1}\mathbf{B}+\mathbf{C}^{-1}\right)^{-1}\mathbf{DA}^{-1} \\ &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BC}\left(\mathbf{I}+\mathbf{DA}^{-1}\mathbf{BC}\right)^{-1}\mathbf{DA}^{-1} \\ &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{I}+\mathbf{CDA}^{-1}\mathbf{B}\right)^{-1}\mathbf{CDA}^{-1}\end{aligned}$$

   Note that the matrices $\mathbf{B}$ and $\mathbf{D}$ need not be invertible for these relationships to be valid. In the latter two, the matrix $\mathbf{C}$ need not be invertible. Notable special cases of this result are

$$\begin{aligned}\left(\mathbf{A}+\mu\mathbf{xy}'\right)^{-1} &= \mathbf{A}^{-1}\left(\mathbf{I} - \frac{\mathbf{xy}'}{\frac{1}{\mu}+\mathbf{y}'\mathbf{A}^{-1}\mathbf{x}}\mathbf{A}^{-1}\right) \\ (\mathbf{A}+\mathbf{B})^{-1} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}\mathbf{A}^{-1}\end{aligned}$$

4. If the matrix $\mathbf{A}$ is either symmetric, Hermitian, circulant, or triangular, its inverse has the same structure. The inverse of a Toeplitz matrix is *not* necessarily Toeplitz. The inverse of a Hankel matrix is symmetric.
5. The inverse of a block diagonal matrix equals the matrix consisting of the inverses of the blocks (if all of these individual inverses exist).

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \cdots \\ \vdots & & \ddots & \vdots \\ \cdots & \cdots & \mathbf{0} & \mathbf{A}_N \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_1^{-1} & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & \mathbf{A}_2^{-1} & \mathbf{0} & \cdots \\ \vdots & & \ddots & \vdots \\ \cdots & \cdots & \mathbf{0} & \mathbf{A}_N^{-1} \end{bmatrix}$$

- A *Givens rotation* is the multiplication of a matrix $\mathbf{A}$ by a unitary matrix $\mathbf{Q}$ that zeroes a specific element of $\mathbf{A}$ [39: pp. 43–7]. This simple operation lies at the core of many numerical algorithms for computing matrix inverses and eigensystems. The rotation matrix has the form of an identity matrix augmented by a square submatrix of sine and cosine rotation values. For example, the matrix that zeroes the second element of the fourth column of a 5×5 matrix has the form

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & c & 0 & s & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -s & 0 & c & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

  where $c = A_{44}/\left(A_{24}^2 + A_{44}^2\right)^{1/2}$ and $s = A_{24}/\left(A_{24}^2 + A_{44}^2\right)^{1/2}$. The product $\mathbf{QA}$ performs the rotation by modifying only the second and fourth rows of $\mathbf{A}$. Column modification is possible by post-multiplying by the rotation matrix and suitably defining the rotation.
- The *determinant* of a square matrix $\mathbf{A}$ is denoted by $\det[\mathbf{A}]$ and is given by

$$\det[\mathbf{A}] = A_{11}\widetilde{A}_{11} + A_{12}\widetilde{A}_{12} + \cdots + A_{1n}\widetilde{A}_{1n},$$

  where $\widetilde{A}_{ij}$, the *cofactor* of $\mathbf{A}$, equals $(-1)^{j+1}$ times the determinant of the $(n-1)\times(n-1)$ matrix formed by deleting the i^{th} row and j^{th} column of $\mathbf{A}$. A non-zero determinant is a necessary and sufficient condition for the existence of the inverse of a matrix. The determinant of $\mathbf{A}'$ equals the conjugate of the determinant of $\mathbf{A}$: $\det[\mathbf{A}'] = \det[\mathbf{A}]^*$. The determinant of a unitary matrix has unity magnitude. The determinant of the product of two square matrices is the product of their determinants.

$$\det[\mathbf{AB}] = \det[\mathbf{A}]\det[\mathbf{B}]$$

The determinant of a sum of matrices is *not* equal to the sum of their determinants. The determinant of a Kronecker product of square matrices equals $\det[\mathbf{A}\otimes\mathbf{B}] = \det[\mathbf{A}]^n \det[\mathbf{B}]^m$, where $\dim[\mathbf{A}] = m$ and $\dim[\mathbf{B}] = n$.

- The *trace* $\mathrm{tr}[\mathbf{A}]$ of the square matrix $\mathbf{A}$ equals the sum of its elements along the main diagonal.

$$\mathrm{tr}[\mathbf{A}] = \sum_{i=1}^{n} A_{ii}$$

The trace of a sum of matrices equals the sum of their traces.

$$\mathrm{tr}[\mathbf{A}+\mathbf{B}] = \mathrm{tr}[\mathbf{A}] + \mathrm{tr}[\mathbf{B}]$$

The trace of a product of two square matrices does *not* necessarily equal the product of the traces of the individual matrices; the product of two matrix's traces does equal the trace of their Kronecker product: $\mathrm{tr}[\mathbf{A}\otimes\mathbf{B}] = \mathrm{tr}[\mathbf{A}]\,\mathrm{tr}[\mathbf{B}]$. However, one of the more interesting properties of the trace is

$$\mathrm{tr}[\mathbf{AB}] = \mathrm{tr}[\mathbf{BA}].$$

Proving this property is straightforward. Using the full expression for the product of two matrices given above, $\mathrm{tr}[\mathbf{AB}] = \sum_i \sum_k A_{ik}B_{ki}$. Writing a similar expression for $\mathbf{BA}$ easily demonstrates this property. With this result, computation of the trace can be simplified in circumstances where the component matrices are not square; whichever product yields the smaller matrix can be used to compute the trace. See the discussion of quadratic forms on page 241 for a *very* useful application of this property.

- The *gradient* with respect to a vector of a scalar-valued function $f(\mathbf{x})$ equals a column matrix of the partial derivatives.

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathrm{col}\left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_N}\right]$$

Examples of gradient calculation derived from this definition are:

$$\begin{aligned}
&\nabla_{\mathbf{x}}\, \mathbf{x}^t\mathbf{y} = \nabla_{\mathbf{x}}\, \mathbf{y}^t\mathbf{x} = \mathbf{y} \\
&\nabla_{\mathbf{x}}\, \mathbf{x}^t\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x} \\
&\nabla_{\mathbf{x}}\, \mathrm{tr}[\mathbf{x}\mathbf{y}^t] = \nabla_{\mathbf{x}}\, \mathrm{tr}[\mathbf{y}\mathbf{x}^t] = \mathbf{y}
\end{aligned}$$

The gradient with respect to a matrix of a scalar-valued function can be defined similarly:

$$[\nabla_{\mathbf{A}} f(\mathbf{A})]_{ij} = \frac{\partial f}{\partial A_{ij}}.$$

Examples are:

$$\begin{aligned}
&\nabla_{\mathbf{A}}\, \mathrm{tr}[\mathbf{A}] = \mathbf{I} \\
&\nabla_{\mathbf{A}}\, \mathrm{tr}[\mathbf{BA}] = \nabla_{\mathbf{A}}\, \mathrm{tr}[\mathbf{AB}] = \mathbf{B}^t && \nabla_{\mathbf{A}}\, \mathrm{tr}[\mathbf{BA}^{-1}] = -[\mathbf{A}^{-1}\mathbf{BA}^{-1}]^t \\
&\nabla_{\mathbf{A}} \exp\{\mathbf{x}^t\mathbf{A}\mathbf{x}\} = \mathbf{x}\mathbf{x}^t \exp\{\mathbf{x}^t\mathbf{A}\mathbf{x}\} && \nabla_{\mathbf{A}}\, \mathbf{x}^t\mathbf{A}^{-1}\mathbf{x} = -\left[\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}\right]^t \\
&\nabla_{\mathbf{A}} \det[\mathbf{A}] = \det[\mathbf{A}]\left(\mathbf{A}^{-1}\right)^t && \nabla_{\mathbf{A}} \ln\det[\mathbf{A}] = \left(\mathbf{A}^{-1}\right)^t
\end{aligned}$$

- The *matrix exponential* if a square matrix $\mathbf{A}$, $\exp[\mathbf{A}]$, is not defined to the element-by-element exponential. Rather, it is defined by the series expansion

$$\exp[\mathbf{A}] = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}$$

The inverse operation $\log[\mathbf{A}]$ is defined to be the *matrix logarithm*. Interestingly, $\det[\exp(\mathbf{A})] = e^{\mathrm{tr}[\mathbf{A}]}$. However, $\exp[\mathbf{A}]\exp[\mathbf{B}]$ *only* equals $\exp[\mathbf{A}+\mathbf{B}]$ if the matrices $\mathbf{A}$, $\mathbf{B}$ commute: $\mathbf{AB} = \mathbf{BA}$.

- The *characteristic polynomial* $p(\lambda)$ of an $n \times n$ square matrix $\mathbf{A}$ is defined to be

$$p(\lambda) = \sum_{k=0}^{n} c_k \lambda^k = \det(\lambda \mathbf{I} - \mathbf{A}) \, .$$

  Note that $c_n = 1$ and that $c_0 = (-1)^n \det \mathbf{A}$. The characteristic polynomial's order is n. The n roots of a matrix's characteristic polynomial are its eigenvalues (see §B.5). The *Cayley-Hamilton Theorem* states that every square matrix satisfies its characteristic polynomial.

$$p(\mathbf{A}) = \sum_{k=0}^{n} c_k \mathbf{A}^k = \mathbf{0}$$

  Because of this result, $\mathbf{A}^n$ equals a matrix polynomial of order $n-1$. Consequently, every matrix function can be written as a matrix polynomial of order $n-1$, in particular the matrix exponential.

B.4 Quadratic Forms

Quadratic forms are key quantities of study in statistical signal processing. They are comprised by a Hermitian matrix $\mathbf{A}$ and a vector $\mathbf{x}$ to produce the real-valued scalar $\mathbf{x}'\mathbf{A}\mathbf{x}$. The matrix is termed the *kernel* of the quadratic form. In the special case that $\mathbf{A} = \mathbf{I}$, the quadratic form reduces to the inner product of $\mathbf{x}$ with itself; quadratic forms are thus considered generalizations of the inner product. If $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all non-zero choices of $\mathbf{x}$, the matrix $\mathbf{A}$ is said to be *positive definite*. If $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ under the same conditions, $\mathbf{A}$ is *non-negative definite*. The structure of Hermitian matrices, even if they are Toeplitz, is not sufficient to guarantee that they are non-negative definite. In contrast, Hankel matrices having no negative elements are always non-negative definite.

The argument of the exponential in the probability density function of a Gaussian random vector is a quadratic form. The kernel of that quadratic form is the covariance matrix of the random vector. Assuming zero mean for algebraic simplicity, the covariance matrix is the expected value of the outer product of the random vector $\mathbf{z}$ with itself: $\mathbf{A} = \mathsf{E}[\mathbf{z}\mathbf{z}']$. Such matrices are *always* non-negative definite. To show this result, simply consider an expanded version of the quadratic form: $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathsf{E}\left[|\mathbf{x}'\mathbf{z}|^2\right]$. Because of the squared magnitude, this quantity can never be negative.

Quadratic forms can be re-written using the trace operation. The trace of a quadratic form simply equals its value because a quadratic form is a scalar. Using the properties of the trace operation,

$$\mathrm{tr}\left[\mathbf{x}'\mathbf{A}\mathbf{x}\right] = \mathrm{tr}\left[\mathbf{A}\mathbf{x}\mathbf{x}'\right] .$$

Thus, a quadratic form equals the trace of the product between the kernel and the outer product of the vector with itself. This seemingly more complicated expression has utility in signal processing. See §2.1.9 {8} on the multivariate Gaussian distribution for the most well-known application of this formula.

Positive-definite Hermitian matrices can be expanded using the *Cholesky factorization*.

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}'$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{L}$ is lower-triangular with all of the diagonal elements equal to unity.

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ L_{21} & 1 & 0 & \cdots \\ L_{31} & L_{32} & \ddots & \ddots \\ \vdots & \vdots & & \ddots \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} D_{11} & 0 & 0 & \cdots \\ 0 & D_{22} & 0 & \cdots \\ \vdots & 0 & D_{33} & \ddots \\ \vdots & \vdots & & \ddots \end{bmatrix}$$

The inverse of a lower-triangular matrix is also lower-triangular; the inverse of the matrix $\mathbf{A}$ is written as $\mathbf{A}^{-1} = \mathbf{L}'^{-1}\mathbf{D}^{-1}\mathbf{L}^{-1}$.

In optimization problems, the quadratic form frequently represents a squared error cost function: the kernel imposes a shape to the range of values of $\mathbf{x}$ and the quadratic form grows quadratically with increasing the vector's length. Analytic solutions to such optimization problems are found by evaluating the *gradient*

of the cost function with respect to the vector $\mathbf{x}$. The gradient with respect to $\mathbf{x}$ is evaluated by treating the vector's conjugate as a constant and *vice versa*. In this way, we find that

$$\nabla_{\mathbf{x}} \mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{A}^t\mathbf{x}^*$$
$$\nabla_{\mathbf{x}^*} \mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}.$$

B.5 Matrix Eigenanalysis

One of the most powerful concepts in matrix algebra is eigenanalysis. Letting the matrix $\mathbf{A}$ be square, the non-zero vector $\mathbf{v}$ is said to be an *eigenvector* of $\mathbf{A}$ if it satisfies

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v},$$

where λ is the scalar termed the *eigenvalue* associated with $\mathbf{v}$. A scaled eigenvector is also an eigenvector. Because of this property, we define an eigenvector to *always* have unit inner product ($\mathbf{v}'\mathbf{v} = 1$). An $n \times n$ matrix has n eigenvalues, which may not be distinct, and n eigenvectors.

The *generalized eigenvector* $\mathbf{v}$ of the matrix pair $(\mathbf{A},\mathbf{B})$ satisfies

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{B}\mathbf{v}.$$

Expressed in a slightly different form, the defining equation of eigenanalysis becomes $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$; for generalized eigenanalysis, $(\mathbf{A} - \lambda\mathbf{B})\mathbf{v} = \mathbf{0}$. If the matrices within the parentheses had inverses, the *only* solutions to these equations would be $\mathbf{v} = \mathbf{0}$, a trivial and uninteresting result. To find any non-zero solutions, the eigenvectors, these matrices must not have inverses, which implies that their determinants must be zero.

$$\det[\mathbf{A} - \lambda\mathbf{I}] = 0 \quad \det[\mathbf{A} - \lambda\mathbf{B}] = 0$$

For an $n \times n$ matrix, each equation becomes an n^{th} order polynomial. The eigenvalues of a matrix are the roots of this polynomial. Because no closed form expression exists for the roots of polynomials of greater than fourth order, the eigenvalues of a matrix must be found numerically in most cases of interest in signal processing. The n eigenvalues are conventionally labelled in decreasing order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Eigenvalues are not necessarily unique; the number of times a value is repeated is termed its multiplicity. For example, every eigenvalue of the identity matrix equals unity and thus has multiplicity n. The eigenvalues of a sum or product of matrices are not easily expressed in terms of the eigenvalues of the component matrices. One remarkably simple result concerns Kronecker products: the eigenvalues of $\mathbf{A} \otimes \mathbf{B}$ equal $\lambda_i^{\mathbf{A}}\lambda_j^{\mathbf{B}}$, $i = 1,\ldots,m$, $j = 1,\ldots,n$. We term the set $\{\lambda_i, \mathbf{v}_i\}_{\mathbf{A}}$ of eigenvectors and associated eigenvalues the *eigensystem* of the matrix $\mathbf{A}$. For generalized systems to have n elements, the matrix $\mathbf{B}$ must be invertible. If not, the number of eigenvectors can be zero, less than n, or infinite [39: p. 252]. When invertible, the generalized eigensystem $\{\lambda_i, \mathbf{v}_i\}_{(\mathbf{A},\mathbf{B})}$ equals the eigensystem $\{\lambda_i, \mathbf{v}_i\}_{\mathbf{B}^{-1}\mathbf{A}}$.

Hermitian Matrices

In the special case where a matrix has Hermitian symmetry, several interesting properties arise. Because of the prevalence of the correlation matrix, which is Hermitian, this situation occurs often in practice.

- *The eigenvalues of a Hermitian matrix are real.* The inner product $\mathbf{v}'\mathbf{v}$ can be expressed as $\mathbf{v}'\mathbf{A}\mathbf{v} = \lambda\mathbf{v}'\mathbf{v}$. The left side of this equation equals its conjugate transpose; since the inner product $\mathbf{v}'\mathbf{v}$ is always real, we have $\lambda^* = \lambda$ and thus the eigenvalues are real.
- *If $\mathbf{A}$ is positive definite, all of its eigenvalues are positive.* The quadratic form $\mathbf{v}'\mathbf{A}\mathbf{v}$ is positive; as inner products are positive, $\mathbf{v}'\mathbf{A}\mathbf{v} = \lambda\mathbf{v}'\mathbf{v}$ implies that λ must also be positive.
- *The eigenvectors associated with distinct eigenvalues of a Hermitian matrix are orthogonal.* Consider $\mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and $\mathbf{A}\mathbf{v}_2 = \lambda_2\mathbf{v}_2$ for $\lambda_1 \neq \lambda_2$. Noting that $\mathbf{v}_2'\mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_2'\mathbf{v}_1$ and $\mathbf{v}_1'\mathbf{A}\mathbf{v}_2 = \lambda_2\mathbf{v}_1'\mathbf{v}_2$, these expressions differ on the left side only in that they are conjugate transposes of each other. Consequently, $(\lambda_2 - \lambda_1)\mathbf{v}_1'\mathbf{v}_2 = 0$, thereby indicating that $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal.

Define a matrix $\mathbf{V}$ having as its columns the eigenvectors of the matrix $\mathbf{A}$.

$$\mathbf{V} = \text{col}[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$$

If $\mathbf{A}$ is Hermitian and has distinct eigenvalues, its eigenvectors are orthogonal, implying that $\mathbf{V}$ must satisfy $\mathbf{V}'\mathbf{V} = \mathbf{I}$ (i.e., $\mathbf{V}$ is unitary). Furthermore, the product $\mathbf{V}'\mathbf{A}\mathbf{V} \equiv \Lambda$ is a diagonal matrix with the eigenvalues lying along the diagonal.

$$\mathbf{V}'\mathbf{A}\mathbf{V} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \equiv \Lambda$$

Because $\mathbf{V}$ is unitary, we find that Hermitian matrices having distinct eigenvalues can be expressed

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}'.$$

This equation defines the matrix's *diagonal form*. From this diagonal form, the determinant and trace of $\mathbf{A}$ are easily related to the eigenvalues of $\mathbf{A}$ as

$$\det[\mathbf{A}] = \prod_{i=1}^{n} \lambda_i$$

$$\text{tr}[\mathbf{A}] = \sum_{i=1}^{n} \lambda_i$$

Expressing the diagonal form of $\mathbf{A}$ less concisely, an important conclusion can be drawn: *any Hermitian matrix* $\mathbf{A}$ *can be expressed with the expansion* $\boxed{\mathbf{A} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i'}$. This result is used frequently when eigenanalysis is applied to signal processing. In particular, quadratic forms can be expressed as

$$\boxed{\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \lambda_i \left|\mathbf{x}'\mathbf{v}_i\right|^2}.$$

In addition, because $\mathbf{v} = \lambda\mathbf{A}^{-1}\mathbf{v}$, the eigenvectors of a matrix and its inverse are identical with the corresponding eigenvalues being reciprocals of each other. Therefore,

$$\mathbf{A}^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i'.$$

For generalized eigenanalysis, orthogonality of eigenvectors is defined with respect to the matrix $\mathbf{B}$: $\mathbf{v}_i'\mathbf{B}\mathbf{v}_j = \delta_{ij}$. Then,*

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{B}\mathbf{v}_i\mathbf{v}_i'\mathbf{B}'$$

$$\mathbf{A}^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i'$$

The matrix exponential† of a Hermitian, not necessarily positive-definite, matrix can be expressed as

$$\exp[\mathbf{A}] = \exp[\mathbf{V}\Lambda\mathbf{V}'] = \mathbf{V}\exp[\Lambda]\mathbf{V}'$$

since the eigenvector matrices are unitary (for example, $\mathbf{V}\Lambda\mathbf{V}' \cdot \mathbf{V}\Lambda\mathbf{V}' = \mathbf{V}\Lambda(\mathbf{V}'\mathbf{V})\Lambda\mathbf{V}' = \mathbf{V}\Lambda^2\mathbf{V}'$). The matrix exponential of a diagonal matrix simply equals a diagonal matrix consisting of the exponential of the diagonal terms. Consequently, the eigenvalues of $\exp[\mathbf{A}]$ are strictly positive. Furthermore, $\mathbf{V}\exp[\Lambda]\mathbf{V}'$ is a symmetric matrix, making $\exp[\mathbf{A}]$ a positive-definite Hermitian matrix. Because the matrix-logarithm is the inverse function of the matrix exponential, a one-to-one correspondence results between symmetric and positive-definite symmetric matrices. A similar correspondence between non-negative definite matrices and symmetric ones does not exist.

*Despite appearances, the second equation is correct. Test its veracity.

†See page 240.

Rayleigh Quotient

An important quantity in matrix algebra easily manipulated in terms of eigenvectors and eigenvalues is the ratio of two quadratic forms.

$$R = \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}}$$

The matrices $\mathbf{A}$ and $\mathbf{B}$ are Hermitian and $\mathbf{B}$ must be positive definite. The key question asked is what vector $\mathbf{x}$ maximizes (or minimizes) this *Rayleigh quotient*. The simplest, and most prevalent case, occurs when $\mathbf{B} = \mathbf{I}$. The Rayleigh quotient is expressed in terms of the eigenvectors and eigenvalues of $\mathbf{A}$.

$$R = \frac{\sum_{i=1}^{n} \lambda_i |\mathbf{x}'\mathbf{v}_i|^2}{\mathbf{x}'\mathbf{x}}$$

The denominator is now seen to normalize the quotient: the quotient's value does not change when $\mathbf{x}$ is multiplied by a scalar. We can thus take the norm of $\mathbf{x}$ to be one and the denominator is no longer a concern. To maximize the numerator, the vector must be proportional to the eigenvector $\mathbf{v}_i$ corresponding to the maximum eigenvalue of $\mathbf{A}$. If not chosen this way, some of $\mathbf{x}$'s components would project onto other eigenvectors which necessarily have smaller eigenvalues and thus provide less weight to the sum. A similar argument applies if the minimizing vector is sought.

$$R_{\min} = \lambda_{\min}, \mathbf{x}_{\min} = \mathbf{v}_{\min} \qquad R_{\max} = \lambda_{\max}, \mathbf{x}_{\max} = \mathbf{v}_{\max}$$

Furthermore, the Rayleigh quotient is bounded by the minimum and maximum eigenvalues.

$$\lambda_{\min} \leq R \leq \lambda_{\max}$$

When $\mathbf{B}$ does not equal the identity matrix, we can use the generalized eigensystem of $(\mathbf{A}, \mathbf{B})$ to find the extrema of the Rayleigh quotient. When $\mathbf{x}$ is a generalized eigenvector, $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$, which implies that the Rayleigh quotient equal λ. Because $\mathbf{A}$ and $\mathbf{B}$ are Hermitian and positive definite, the generalized eigenvectors form a basis and the Rayleigh quotient obeys the above inequality for the generalized eigenvalues. We can express the Rayleigh quotient in other ways. First of all, we know that the generalized eigensystem equals the eigensystem of $\mathbf{B}^{-1}\mathbf{A}$. In addition, define $\tilde{\mathbf{x}} = \mathbf{B}^{1/2}\mathbf{x}$ where $\mathbf{B}^{1/2}$ is the square root of $\mathbf{B}$. Computationally, square roots may seem difficult to find; in eigenanalysis terms, a straightforward definition emerges.

$$\mathbf{B}^{1/2} \equiv \sum_{i=1}^{n} \lambda_i^{1/2} \mathbf{v}_i \mathbf{v}_i'$$

With this definition, the square root of a matrix satisfies the intuitive property $\mathbf{B}^{1/2}\mathbf{B}^{1/2} = \mathbf{B}$. The Rayleigh quotient thus becomes

$$R = \frac{\tilde{\mathbf{x}}' \left(\mathbf{B}^{-1/2}\right)' \mathbf{A} \left(\mathbf{B}^{-1/2}\right) \tilde{\mathbf{x}}}{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}}$$

Thus, the vector $\tilde{\mathbf{x}}$ maximizing or minimizing the Rayleigh quotient corresponds to the eigenvector having maximum or minimum eigenvalues of the matrix product $\left(\mathbf{B}^{-1/2}\right)' \mathbf{A} \left(\mathbf{B}^{-1/2}\right)$.

Singular Value Decomposition

Related to eigenanalysis is the *singular value decomposition* or *SVD* technique. Letting $\mathbf{A}$ be an $m \times n$ matrix consisting of complex entries, it can be expressed by

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}',$$

where $\mathbf{D}$ is the $k \times k$ diagonal matrix $\text{diag}[\sigma_1, \ldots, \sigma_k]$ and where $\mathbf{U}$ $(m \times k)$ and $\mathbf{V}$ $(k \times n)$ are matrices satisfying $\mathbf{U}'\mathbf{U} = \mathbf{I}_k$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}_k$. Letting $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the i^{th} columns of $\mathbf{U}$ and $\mathbf{V}$ respectively, then

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i \quad \mathbf{u}_i'\mathbf{A} = \sigma_i \mathbf{v}_i'$$

The scalars σ_i are termed the singular values of the matrix $\mathbf{A}$ while $\mathbf{u}_i$ and $\mathbf{v}_i$ are termed the left and right (respectively) eigenvectors of $\mathbf{A}$. The number of non-zero singular values equals k, which must not exceed $\min(m,n)$. The number equals this upper limit when the side of the matrix having the smallest length (either the rows or the columns) has linearly independent components. From these expressions, the eigenvectors of $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ are found to be

$$\mathbf{A}'\mathbf{A}\mathbf{v}_i = \sigma_i^2\mathbf{v}_i \quad \mathbf{A}\mathbf{A}'\mathbf{u}_i = \sigma_i^2\mathbf{u}_i.$$

One of these products will not have full rank when the matrix $\mathbf{A}$ is not square. For example, if $\mathbf{A}$ is a column vector, $\mathbf{A}'\mathbf{A}$ is a scalar and thus is invertible while $\mathbf{A}\mathbf{A}'$ is an $m \times m$ matrix having only one non-zero singular value. These matrices are Hermitian; they share the same non-zero eigenvalues and these equal the *squares* of the singular values. Furthermore, the collections of vectors $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_i\}$ are each orthonormal sets. This property means that $\mathbf{U}\mathbf{U}' = \mathbf{I}_m$ and $\mathbf{V}\mathbf{V}' = \mathbf{I}_n$. The singular value decomposition suggests that all matrices have an orthonormal-like expansion of the form

$$\boxed{\mathbf{A} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i'}.$$

No non-square matrix has an inverse. Singular value decomposition can be used to define the *pseudo inverse* of a rectangular matrix. Assuming that all of the singular values are non-zero, the pseudo inverse $\mathbf{A}^{\sim 1}$ satisfies either $\mathbf{A}\mathbf{A}^{\sim 1} = \mathbf{I}_m$ or $\mathbf{A}^{\sim 1}\mathbf{A} = \mathbf{I}_n$ according to which dimension of the matrix is the smallest. By analogy to the eigenvalue-eigenvector expansion of an invertible matrix, the singular value decomposition of a matrix's pseudo inverse would be

$$\mathbf{A}^{\sim 1} = \sum_{i=1}^{k} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i'$$

or in matrix terms $\mathbf{A}^{\sim 1} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}'$. The pseudo inverses can be defined more directly by either $\mathbf{A}^{\sim 1} = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}$ or $\mathbf{A}^{\sim 1} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$. To be more concrete, suppose $\mathbf{A}$ has linearly independent rows. Its pseudo inverse is given by $\mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}$ so that $\mathbf{A}\mathbf{A}^{\sim 1} = \mathbf{I}$: only the right inverse exists.

B.6 Projection Matrices

Eigenanalysis and singular value analysis are often used because they express a matrix's properties "naturally". One case in point is the projection matrix. A projection matrix $\mathbf{P}$ has the property $\mathbf{P}^m = \mathbf{P}^n$ for $m,n \geq 0$. The eigenvalues of a projection must therefore obey the relationship $\lambda^m = \lambda^n$. Two solutions are possible: $\lambda = 1$ and $\lambda = 0$. Thus, the eigenvectors of a projection matrix either have zero eigenvalues, in which case any vector that can be expressed by them is annihilated by the matrix ($\mathbf{P}\mathbf{x} = \mathbf{0}$), or the eigenvectors have unity eigenvalues and vectors comprised of them are unaffected by the matrix ($\mathbf{P}\mathbf{x} = \mathbf{x}$). If a vector has components belonging to both sets of eigenvectors, the part of the vector that "survives" the matrix is that portion represented by the eigenvectors having unity eigenvalues. Hence the origin of the name "projection matrix": matrices having only unity and zero eigenvalues can be used to find those components of any vector which correspond to the eigenvectors having unity eigenvalues.

For any $m \times n$ matrix $\mathbf{A}$ ($m < n$) that is used as a linear transformation, it defines a subspace for the *domain* of the transformation. In more exact terms, when the m-dimensional vectors $\mathbf{y} = \mathbf{A}\mathbf{x}$ define a vector space, what is the image of this space in the original, higher dimensional space containing the vectors $\mathbf{x}$? The most elegant response is in terms of the singular values and singular vectors of $\mathbf{A}$. The vector $\mathbf{y}$ is thereby expressed

$$\mathbf{y} = \sum_{i=1}^{m} \sigma_i (\mathbf{v}_i'\mathbf{x}) \mathbf{u}_i \,.$$

The result of the linear transformation is always a linear combination of the left singular vectors of $\mathbf{A}$.

The projection matrix $\mathbf{P}_\mathbf{A}$ associated with a linear transformation $\mathbf{A}$ would project all n-dimensional vectors onto a subspace defined by the linear transformation. $\mathbf{P}_\mathbf{A}$ would have m unity eigenvalues for the left

singular vectors of $\mathbf{A}$ and zero eigenvalues for all remaining $m-n$ eigenvectors. Defining $\mathbf{U}^{\perp}$ to be the matrix of eigenvectors corresponding to these zero eigenvalues, the projection matrix can be expressed as

$$\mathbf{P}_{\mathbf{A}} = \begin{bmatrix} \mathbf{U} & \mathbf{U}^{\perp} \end{bmatrix} \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U} & \mathbf{U}^{\perp} \end{bmatrix}'.$$

This projection matrix can be expressed more directly in terms of the matrix $\mathbf{A}$ as $\mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}\mathbf{A}$. Note than this expression can be written in terms of the transformation's pseudo inverse: while $\mathbf{A}\mathbf{A}^{\sim 1} = \mathbf{I}_m$, $\mathbf{P}_{\mathbf{A}} = \mathbf{A}^{\sim 1}\mathbf{A}$.

Appendix C

Ali-Silvey Distances

Ali-Silvey distances comprise a family of quantities that depend on the likelihood ratio $\Lambda(\mathbf{X})$ and on the model-describing densities p_0, p_1 in the following way.

$$d(p_0, p_1) = f(\mathsf{E}_0[c(\Lambda(\mathbf{X}))])$$

Here, $f(\cdot)$ is a non-decreasing function, $c(\cdot)$ a convex function, and $\mathsf{E}_0[\cdot]$ means expectation with respect to p_0. Where applicable, π_0, π_1 denote the *a priori* probabilities of the models. Basseville [8] is good reference on distances in this class and many others. In all cases, the observations consist of L IID random variables.

Ali-Silvey Distances and Relation to Detection Performance			
Name	$c(\cdot)$	Performance	Comment
Kullback-Leibler $\mathscr{D}(p_1 \| p_0)$	$(\cdot)\log(\cdot)$	$\lim_{L\to\infty} -\frac{1}{L}\log P_F = d(p_0, p_1)$	Neyman-Pearson error rate under both fixed and exponentially decaying constraints on P_M (P_F)
Kullback-Leibler $\mathscr{D}(p_0 \| p_1)$	$-\log(\cdot)$	$\lim_{L\to\infty} -\frac{1}{L}\log P_M = d(p_0, p_1)$	
J-Divergence	$((\cdot)-1)\log(\cdot)$	$\pi_0\pi_1 \exp\{-d(p_0, p_1)/2\} \leq P_e$	$J(p_0, p_1) = \mathscr{D}(p_0 \| p_1) + \mathscr{D}(p_1 \| p_0)$
Chernoff	$(\cdot)^s, s \in (0,1)$	$\max \lim_{L\to\infty} -\frac{1}{L}\log P_e = \inf_{s\in(0,1)} d(p_0, p_1)$	Independent of *a priori* probabilities Not in Ali-Silvey class
M-Hypothesis Chernoff	$(\cdot)^s, s \in (0,1)$	$\max \lim_{L\to\infty} -\frac{1}{L}\log P_e = \min_{i\neq j} \inf_s d(p_i, p_j)$	
Bhattacharyya	$(\cdot)^{1/2}$	$\pi_0\pi_1[d(p_0, p_1)]^2 \leq P_e \leq \sqrt{\pi_0\pi_1}\, d(p_0, p_1)$	Minimizing $d(p_0, p_1)$ will tend to minimize P_e
Orsak	$\lvert\pi_1(\cdot) - \pi_0\rvert$	$P_e = \frac{1}{2} - \frac{1}{2}d(p_0, p_1)$	Exact formula for average error probability
Kolmogorov	$\frac{1}{2}\lvert(\cdot) - 1\rvert$	If $\pi_0 = \pi_1$, $P_e = \frac{1}{2} - \frac{1}{2}d(p_0, p_1)$	
Hellinger	$((\cdot)^{1/2} - 1)^2$		

Bibliography

1. M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. U.S. Government Printing Office, 1968.
2. H. Akaike. A new look at the statistical model identification problem. *IEEE Trans. Auto. Control*, AC-19:716–723, Dec. 1974.
3. S. T. Alexander. *Adaptive Signal Processing*. Springer-Verlag, New York, 1986.
4. B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice Hall, Englewood Cliffs, NJ, 1979.
5. A. Antoniadis, J. Bigot, and T. Sapatinas. Wavelet estimators in nonparametric regression: a comparative simulation study. *J. Statistical Software*, 6:1–83, 2001.
6. A. B. Baggeroer. Confidence intervals for regression (MEM) spectral estimates. *IEEE Trans. Info. Th.*, IT-22:534–545, Sept. 1976.
7. M. S. Bartlett. Smoothing periodograms from time series with continuous spectra. *Nature*, 161:686–687, 1948.
8. M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18:349–369, 1989.
9. G. J. Bierman. *Factorization Methods for Discrete Sequential Estimation*. Academic Press, New York, 1977.
10. R. B. Blackman and J. W. Tukey. *The Measurement of Power Spectra*. Dover Press, New York, 1958.
11. H. W. Bode and C. E. Shannon. A simplified derivation of linear least-squares smoothing and prediction theory. *Proc. IRE*, 38:417–426, 1950.
12. J. P. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, Department of Geophysics, Stanford University, Stanford, CA, 1975.
13. J. P. Burg. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, 37:375–376, Apr. 1972.
14. C.S. Burrus, R.A. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice-Hall, 1988.
15. D. Burshtein and E. Weinstein. Confidence intervals for the maximum entropy spectrum. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-35:504–510, Apr. 1987.
16. T. Cai and B. Silverman. Incorporate information on neighboring coefficients into wavelet estimation. *Indian J. Statistics*, 63:127–148, 2001.
17. J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE*, 57:1408–1418, 1969.
18. J. Capon. Maximum-likelihood spectral estimation. In S. Haykin, editor, *Nonlinear Methods of Spectral Analysis*, pages 155–179. Springer-Verlag, New York, 1979.
19. J. Capon and N. R. Goodman. Probability distributions for estimators of the frequency-wavenumber spectrum. *Proc. IEEE*, 58:1785–1786, Oct. 1970.
20. J. W. Carlyle and J. B. Thomas. On nonparametric signal detectors. *IEEE Trans. Info. Th.*, IT-10:146–152, Apr. 1964.
21. D. Chazan, M. Zakai, and J. Ziv. Improved lower bounds on signal parameter estimation. *IEEE Trans. Info. Th.*, IT-21:90–93, Jan. 1975.

22. H. Chernoff. Measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.*, 23:493–507, 1952.
23. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., second edition, 2006.
24. C. F. N. Cowan and P. M. Grant, editors. *Adaptive Filters*. Prentice Hall, Englewood Cliffs, NJ, 1985.
25. H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946.
26. H. Cramér. *Random Variables and Probability Distributions*. Cambridge University Press, third edition, 1970.
27. J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, Englewood Cliffs, NJ, 1983.
28. D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *J. R. Statist. Soc., Series B*, 57(2):301–337, 1995.
29. D.L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1:100–115, 1993.
30. D.L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. American Statistical Association*, 90:1200–1224, 1995.
31. D. J. Edelblute, J. M. Fisk, and G. L. Kinnison. Criteria for optimum-signal-detection theory for arrays. *J. Acoust. Soc. Am.*, 41:199–205, Jan. 1967.
32. B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.
33. A. H. El-Sawy and V. D. Vandelinde. Robust detection of known signals. *IEEE Trans. Info. Th.*, IT-23:722–727, 1977.
34. R.M. Fagan. Information measures: Statistical confidence limits and inference. *J. Th. Biol.*, 73:61–79, 1978.
35. A. Feuer and E. Weinstein. Convergence analysis of LMS filters with uncorrelated Gaussian data. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-33:222–229, Feb. 1985.
36. B. Friedlander. Adaptive algorithms for finite impulse response filters. In Cowan and Grant [24], pages 29–59.
37. A. Gelb, editor. *Applied Optimal Estimation*. MIT Press, Cambridge, MA, 1974.
38. J. D. Gibson and J. L. Melsa. *Introduction to Non-Parametric Detection with Applications*. Academic Press, New York, 1975.
39. G. H. Golub and C. F. van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1983.
40. A. Grossmann and J. Morlet. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, 15(4):723–736, 1984.
41. M. Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. Info. Theory*, 35:401–408, 1989.
42. A. Haar. Zür Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69:331–371, 1910.
43. P. Hall. *Rates of convergence in the central limit theorem*, volume 62 of *Research Notes in Mathematics*. Pitman Advanced Publishing Program, 1982.
44. F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*, 66:51–83, Jan. 1978.
45. S. Haykin. *Adaptive Filter Theory*. Prentice Hall, Englewood Cliffs, NJ, 1986.
46. C. W. Helstrom. *Statistical Theory of Signal Detection*. Pergamon Press, Oxford, second edition, 1968.
47. M. L. Honig and D. G. Messerschmitt. *Adaptive Filters*. Kluwer Academic Publishers, Boston, 1984.
48. H. Huang and N. Cressie. Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics*, 42:262–276, 2000.
49. P. J. Huber. A robust version of the probability ratio test. *Ann. Math. Stat.*, 36:1753–1758, 1965.
50. P. J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.
51. E. T. Jaynes. On the rationale of the maximum entropy method. *Proc. IEEE*, 70:939–952, Sept. 1982.

52. G. M. Jenkins and D. G. Watts. *Spectral Analysis and its Applications*. Holden-Day, Inc., San Francisco, 1968.
53. D. H. Johnson. The application of spectral estimation methods to bearing estimation problems. *Proc. IEEE*, 70:1018–1028, Sept. 1982.
54. D. H. Johnson and G. C. Orsak. Relation of signal set choice to the performance of optimal non-Gaussian detectors. *IEEE Trans. Comm.*, 41:1319–1328, 1993.
55. R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Engineering* (ASME Trans.), 82D:35–45, 1960.
56. R. L. Kashyap. Inconsistency of the AIC rule for estimating the order of autoregressive models. *IEEE Trans. Auto. Control*, AC-25:996–998, Oct. 1980.
57. S. A. Kassam. *Signal Detection in Non-Gaussian Noise*. Springer-Verlag, New York, 1988.
58. S. A. Kassam and H. V. Poor. Robust techniques for signal processing: A survey. *Proc. IEEE*, 73:433–481, 1985.
59. S. A. Kassam and J. B. Thomas, editors. *Nonparametric Detection: Theory and Applications*. Dowden, Hutchinson & Ross, Stroudsburg, PA, 1980.
60. S. M. Kay. *Modern Spectral Estimation*. Prentice Hall, Englewood Cliffs, NJ, 1988.
61. E. J. Kelly, I. S. Reed, and W. L. Root. The detection of radar echoes in noise. I. *J. Soc. Indust. Appl. Math.*, 8:309–341, June 1960.
62. E. J. Kelly, I. S. Reed, and W. L. Root. The detection of radar echoes in noise. II. *J. Soc. Indust. Appl. Math.*, 8:481–507, Sept. 1960.
63. L. H. Koopmans. *The Spectral Analysis of Time Series*. Academic Press, New York, 1974.
64. R. T. Lacoss. Data adaptive spectral analysis methods. *Geophysics*, 36:661–675, Aug. 1971.
65. E. L. Lehmann. *Testing Statistical Hypotheses*. John Wiley & Sons, New York, second edition, 1986.
66. R. S. Lipster and A. N. Shiryayev. *Statistics of Random Processes I: General Theory*. Springer-Verlag, New York, 1977.
67. L. Ljung, M. Morf, and D. Falconer. Fast calculation of gain matrices for recursive estimation techniques. *Inter. J. Control*, 27:1–19, Jan. 1978.
68. F. W. Machell and C. S. Penrod. Probability density functions of ocean acoustic noise processes. In E. J. Wegman and J. G. Smith, editors, *Statistical Signal Processing*, pages 211–221. Marcel Dekker, New York, 1984.
69. J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63:561–580, Apr. 1975.
70. J. D. Markel and A. H. Gray, Jr. *Linear Prediction of Speech*. Springer-Verlag, New York, 1976.
71. S. L. Marple, Jr. *Digital Spectral Analysis*. Prentice Hall, Englewood Cliffs, NJ, 1987.
72. D. P. McGinn and D. H. Johnson. Estimation of all-pole model parameters from noise-corrupted sequences. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-37:433–436, Mar. 1989.
73. D. Middleton. Statistical-physical models of electromagnetic interference. *IEEE Trans. Electromag. Compat.*, EMC-17:106–127, 1977.
74. A. R. Milne and J. H. Ganton. Ambient noise under Arctic sea ice. *J. Acoust. Soc. Am.*, 36:855–863, 1964.
75. B. R. Musicus. Fast MLM power spectrum estimation from uniformly spaced correlations. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-33:1333–1335, Oct. 1985.
76. J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. Ser. A*, 231:289–337, Feb. 1933.
77. A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1989.
78. A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, second edition, 1984.
79. E. Parzen. *Stochastic Processes*. Holden-Day, San Francisco, 1962.
80. H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, New York, 1988.
81. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, Sept. 1978.

82. H. Sakai and H. Tokumaru. Statistical analysis of a spectral estimator for ARMA processes. *IEEE Trans. Auto. Control*, AC-25:122–124, Feb. 1980.
83. G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, Mar. 1978.
84. B. W. Silverman. *Density Estimation*. Chapman & Hall, London, 1986.
85. D. L. Snyder. *Random Point Processes*. Wiley, New York, 1975.
86. A. D. Spaulding. Locally optimum and suboptimum detector performance in a non-Gaussian interference environment. *IEEE Trans. Comm.*, COM-33:509–517, 1985.
87. A. D. Spaulding and D. Middleton. Optimum reception in an impulsive interference environment—Part I: Coherent detection. *IEEE Trans. Comm.*, COM-25:910–923, 1977.
88. J. R. Thompson and R. A. Tapia. *Nonparametric Function Estimation, Modeling, and Simulation*. SIAM, Philadelphia, PA, 1990.
89. H. L. van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley & Sons, New York, 1968.
90. A. Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947.
91. M. T. Wasan. *Stochastic Approximation*. Cambridge University Press, 1969.
92. E. Weinstein and A. J. Weiss. A general class of lower bounds in parameter estimation. *ittrans*, 34(2):338–342, March 1988.
93. A. J. Weiss and E. Weinstein. Fundamental limitations in passive time delay estimation: I. Narrow-band systems. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-31:472–486, Apr. 1983.
94. P. D. Welch. The use of the fast Fourier transform for the estimation of power spectra: A method based on time averaging over short modified periodograms. *IEEE Trans. Audio Electroacoust.*, AU-15:70–73, Mar. 1967.
95. B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1985.
96. N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, Cambridge, MA, 1949.
97. R. P. Wishner. Distribution of the normalized periodogram detector. *IRE Trans. Info. Th.*, IT-8:342–349, Sept. 1962.
98. P. M. Woodward. *Probability and Information Theory, with Applications to Radar*. Pergamon Press, Oxford, second edition, 1964.
99. J. Ziv and M. Zakai. Some lower bounds on signal parameter estimation. *IEEE Trans. Info. Th.*, IT-15:386–391, May 1969.

Index