Signal Detection and Classification

Alfred Hero
University of Michigan, Ann Arbor

1 Introduction

Detection and classification arise in signal processing problems whenever a decision is to be made among a finite number of hypotheses concerning an observed waveform. Signal detection algorithms decide whether the waveform consists of "noise alone" or "signal masked by noise." Signal classification algorithms decide whether a detected signal belongs to one or another of prespecified classes of signals. The objective of signal detection and classification theory is to specify systematic strategies for designing algorithms which minimize the average number of decision errors. This theory is grounded in the mathematical discipline of statistical decision theory where detection and classification are respectively called binary and M-ary *hypothesis testing* [1, 2]. However, signal processing engineers must also contend with the exceedingly large size of signal processing datasets, the absence of reliable and tractible signal models, the associated requirement of fast algorithms, and the requirement for real time imbedding of unsupervised algorithms into specialized software or hardware. While ad hoc statistical detection algorithms were implemented by engineers before 1950, the systematic development of signal detection theory was first undertaken by radar and radio engineers in the early 1950's [3],[4].

This chapter provides a brief and limited overview of some of the theory and practice of signal detection and classification. The focus will be on the Gaussian observation model. For more details and examples see the cited references.

2 Signal Detection

Assume that for some physical measurement a sensor produces an output waveform $x = \{x(t) : t \in [0,T]\}$ over a time interval $[0,T]$. Assume that the waveform may have been produced by ambient noise alone or by an impinging signal of known form plus the noise. These two possibilities are called the *null hypothesis* H and the *alternative hypothesis* K, respectively, and are commonly written in the compact notation:

$$\begin{aligned} H &: x = \text{noise alone} \\ K &: x = \text{signal} + \text{noise}. \end{aligned}$$

The hypotheses H and K are called *simple hypotheses* when the statistical distributions of x under H and K involve no unknown parameters such as signal amplitude, signal phase, or noise power. When the statistical distribution of x under a hypothesis depends on unknown (*nuisance*) parameters the hypothesis is called a *composite hypothesis.*

To decide between the null and alternative hypotheses one might apply a high threshold to the sensor output x and make a decision that the signal is present if and only if the threshold is exceeded at some time within $[0,T]$. The engineer is then faced with practical question of where to set the threshold so as to ensure that the number of decision errors is small. There are two types of error possible: the error of missing the signal (decide H under K (signal is present)) and the error of false alarm (decide K under H (no signal is present)). There is always a compromise between choosing a high threshold to make the average number of false alarms small versus choosing a low threshold to make the average number of misses small. To quantify this compromise it becomes necessary to specify the statistical distribution of x under each of the hypotheses H and K.

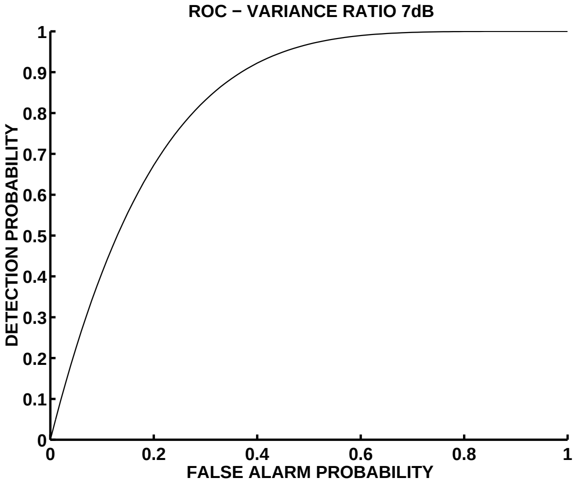

Figure 1: The receiver operating characteristic (ROC) curve describes the tradeoff between maximizing the power P_D and minimizing the probability of false alarm P_{FA} of a test between two hypotheses H and K. Shown is the ROC curve of the LRT (energy detector) which tests between H : x = *complex random variable with variance* $\sigma^2 = 1$, versus K : x = *complex random variable with variance* $\sigma^2 = 5$ (7dB variance ratio).

2.1 The ROC Curve

Let the aforementioned threshold be denoted γ. Define the K decision region $\mathcal{R}_K = \{x : x(t) > \gamma, \quad \text{for some } t \in [0,T]\}$. This region is also called the *critical region* and simply specifies the conditions on x for which the detector declares the signal to be present. Since the detectors makes mutually exclusive binary decisions the critical region completely specifies the operation of the detector. The probabilities of false alarm and miss are functions of γ given by $P_{FA} = P(\mathcal{R}_K|H)$ and $P_M = 1 - P(\mathcal{R}_K|K)$ where $P(A|H)$ and $P(A|K)$ denote the probabilities of arbitrary event A under hypothesis H and hypothesis K, respectively. The probability of correct detection $P_D = P(\mathcal{R}_K|K)$ is commonly called the *power* of the detector and P_{FA} is called the *level* of the detector.

The plot of the pair $P_{FA} = P_{FA}(\gamma)$ and $P_D = P_D(\gamma)$ over the range of thresholds $-\infty < \gamma < \infty$ produces a curve called the receiver operating characteristic (ROC) which completely describes the error rate of the detector as a function of γ (Fig. 1). Good detectors have ROC curves which have desirable properties such as concavity (negative curvature), monotone increase in P_D as P_{FA} increases, high slope of P_D at the point $(P_{FA}, P_D) = (0,0)$, etc. [5]. For the energy detection example shown in Fig. 1 it is evident that regardless of the actual energy σ^2 an increase in the rate of correct detections P_D can be bought only at the expense of increasing the rate of false alarms

P_{FA}. Simply stated, the job of the signal processing engineer is to find ways to test between K and H which push the ROC curve towards the upper left corner of Fig. 1 where P_D is high for low P_{FA}: this is the regime of P_D and P_{FA} where reliable signal detection can occur.

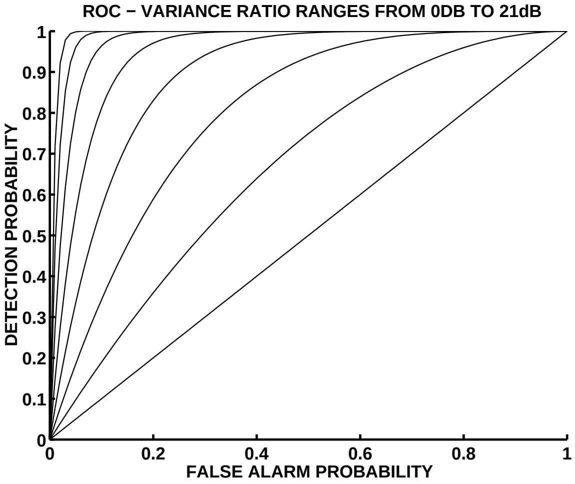

Figure 2: Eight members of the family of ROC curves for the LRT (energy detector) which tests between $H : x =$ *complex random variable with variance* $\sigma^2 = 1$, versus composite $K : x =$ *complex random variable with variance* $\sigma^2 > 1$. ROC curves shown are indexed over a range [0dB, 21dB] of variance ratios in equal 3dB increments. ROC curves approach a step function as variance ratio increases.

2.2 Detector Design Strategies

When the signal waveform and the noise statistics are fully known the hypotheses are simple and an optimal detector exists which has a ROC curve that upper bounds the ROC of any other detector, i.e. it has the highest possible power P_D for any fixed level P_{FA}. This optimal detector is called the most powerful (MP) test and is specified by the ubiquitous likelihood ratio test described below. In the more common case where the signal and/or noise are described by unknown parameters, at least one hypothesis is composite and a detector has different ROC curves for different values of the parameters (see Figure 2). Unfortunately, there seldom exists a uniformly most powerful detector whose ROC curves remain upper bounds for the entire range of unknown parameters. Therefore, for composite hypotheses other design strategies must generally be adopted to ensure reliable detection performance. There are a wide range of different strategies available including: Bayesian detection [5] and hypothesis testing [6], min-max hypothesis testing [2], CFAR detection [7] and similar, unbiased hypothesis testing [1], invariant hypothesis testing [8, 9], sequential detection [10], simultaneous detection and estimation [11], and non-parametric detection [12]. Detailed discussion of these strategies is outside of the scope of this chapter. However, all of these strategies have a common link: their application produces one form or another of the *likelihood ratio test.*

2.3 Likelihood Ratio Test

Here we introduce an unknown parameter θ to simplify the upcoming discussion on composite hypothesis testing. Define the probability density of the measurement x as $f(x|\theta)$ where θ belongs to a parameter space Θ. It is assumed that $f(x|\theta)$ is a known function of x and θ. We can now state the detection problem as the problem of testing between

$$H \; : \; x \sim f(x|\theta), \;\; \theta \in \Theta_H \qquad (1)$$

$$K \; : \; x \sim f(x|\theta), \;\; \theta \in \Theta_K, \qquad (2)$$

where Θ_H and Θ_K are non-empty sets which partition the parameter space into two regions. Note it is essential that Θ_H and Θ_K be *disjoint* ($\Theta_H \cap \Theta_K = \emptyset$) so as to remove any ambiguity on the decisions, and *exhaustive* ($\Theta_H \cup \Theta_K = \Theta$) to ensure that all states of nature in Θ are accounted for. Let a detector be specified by a critical region $\mathcal{R}_K$. Then for any pair of parameters $\theta_H \in \Theta_H$ and $\theta_K \in \Theta_K$ the level and power of the detector can be computed by integrating the probability density $f(x|\theta)$ over $\mathcal{R}_K$

$$P_{FA} = \int_{x \in \mathcal{R}_K} f(x|\theta_H)dx, \qquad (3)$$

and

$$P_D = \int_{x \in \mathcal{R}_K} f(x|\theta_K)dx. \qquad (4)$$

The hypotheses (1) and (2) are simple when $\Theta = \{\theta_H, \theta_K\}$ consists of only two values and $\Theta_H = \{\theta_H\}$ and $\Theta_K = \{\theta_K\}$ are point sets. For simple hypotheses the Neyman-Pearson Lemma [1] states that there exists a most powerful test which maximizes P_D subject to the constraint that $P_{FA} \leq \alpha$, where α is a prespecified maximum level of false alarm. This test has the form of a threshold test known as the likelihood ratio test (LRT)

$$L(x) \stackrel{\text{def}}{=} \frac{f(x|\theta_K)}{f(x|\theta_H)} \underset{H}{\overset{K}{\gtrless}} \eta, \qquad (5)$$

where η is a threshold which is determined by the constraint $P_{FA} = \alpha$

$$\int_{\eta}^{\infty} g(l|\theta_H)dl = \alpha. \qquad (6)$$

Here $g(l|\theta)$ is the probability density function of the likelihood ratio statistic $L(x)$. It must also be mentioned that if the density $g(l|\theta_H)$ contains delta functions a simple randomization [1] of the LRT may be required to meet the false alarm constraint (6).

The test statistic $L(x)$ is a measure of the strength of the evidence provided by x that the probability density $f(x|\theta_K)$ produced x as opposed to the probability density $f(x|\theta_H)$. Similarly, the threshold η represents the detector designer's prior level of "reasonable doubt" about the sufficiency of the evidence - only above a level η is the evidence sufficient for rejecting H.

When θ takes on more than two values at least one of the hypotheses (1) or (2) are composite the Neyman Pearson lemma no longer applies. A popular but ad hoc alternative which enjoys some asymptotic optimality properties is to implement the *generalized likelihood ratio test* (GLRT):

$$L_g(x) \stackrel{\text{def}}{=} \frac{\max_{\theta_K \in \Theta_K} f(x|\theta_K)}{\max_{\theta_H \in \Theta_H} f(x|\theta_H)} \underset{H}{\overset{K}{\gtrless}} \eta \tag{7}$$

where, if possible, the threshold η is set to attain a specified level of P_{FA}. The GLRT can be interpreted as a LRT which is based on the most likely values of the unknown parameters θ_H and θ_K, i.e. the values which maximize the *likelihood functions* $f(x|\theta_H)$ and $f(x|\theta_K)$, respectively (See section on parameter estimation - this chapter).

3 Signal Classification

When, based on a noisy observed waveform x, one must decide among a number of possible signal waveforms $s_1, \ldots, s_p$, $p > 1$, we have a *p-ary signal classification problem.* Denoting $f(x|\theta_i)$ the density function of x when signal s_i is present, the classification problem can be stated as the problem of testing between the p hypotheses

$$\begin{array}{ccl} H_1 & : & x \sim f(x|\theta_1), \theta_1 \in \Theta_1 \\ \vdots & \vdots & \vdots \\ H_p & : & x \sim f(x|\theta_p), \theta_p \in \Theta_p \end{array}$$

where Θ_i is a space of unknowns which parameterize the signal s_i. As before, it is essential that the hypotheses be disjoint, which ensures that $\{f(x|\theta_i)\}_{i=1}^p$ are distinct functions of x for all $\theta_i \in \Theta_i$, $i = 1, \ldots, p$, and that they be exhaustive, which ensures that the true density of x is included in one of the hypotheses. Similarly to the case of detection, a classifier is specified by a partition of the space of observations x into p disjoint decision regions $\mathcal{R}_{H_1}, \ldots, \mathcal{R}_{H_p}$. Only $p-1$ of these decision regions are needed to specify the operation of the classifier. The performance of a signal classifier is characterized by its set of p *misclassification* probabilities $P_{M_1} = 1 - P(x \in \mathcal{R}_{H_1}|H_1), \ldots, P_{M_p} = P(x \in \mathcal{R}_{H_p}|H_p)$. Unlike in the case of detection, even for simple hypotheses, where $\Theta_i = \{\theta_i\}$ consists of a single point, $i = 1, \ldots, p$, optimal p-ary classifiers that uniformly minimize all P_{M_i}'s do not exist for $p > 2$. However classifiers can be designed to minimize other weaker criteria such as average misclassification probability $\frac{1}{p}\sum_{i=1}^p P_{M_i}$ [5], worst case misclassification probability $\max_i P_{M_i}$ [2], Bayes posterior misclassification probability [13], and others.

The maximum likelihood (ML) classifier is a popular classification technique which is closely related to maximum likelihood parameter estimation. This classifier is specified by the rule

$$\text{decide } H_j \text{ if and only if } \max_{\theta_j \in \Theta_j} f(x|\theta_j) \geq \max_k \max_{\theta_k \in \Theta_k} f(x|\theta_k), \quad j = 1, \ldots, p. \tag{8}$$

When the signal waveforms and noise statistics subsumed by the hypotheses $H_1, \ldots, H_p$ are fully known the ML classifier takes the simpler form:

$$\text{decide } H_j \text{ if and only if } f_j(x) \geq \max_k f_k(x), \quad j = 1, \ldots, p$$

where f_k denotes the known density function of x when the k-th signal is present. For this simple case it can be shown that the ML classifier is an optimal decision rule which minimizes the total misclassification error probability, as measured by the average $\frac{1}{p}\sum_{i=1}^{p} P_{M_i}$. In some cases a weighted average $\frac{1}{p}\sum_{i=1}^{p} \beta_i P_{M_i}$ is a more appropriate measure of total misclassification error, e.g. when β_i is the prior probability of H_i, $i = 1, \ldots, p$, $\sum_{i=1}^{p} \beta_i = 1$. For this case, the optimal classifier is given by the *maximum a posteriori* (MAP) decision rule [13, 5]

$$\text{decide } H_j \text{ if and only if } f_j(x)\beta_j \geq \max_k f_k(x)\beta_k, \quad j = 1, \ldots, p.$$

4 The Linear Multivariate Gaussian Model

Assume that $\mathbf{X}$ is an $m \times n$ matrix of complex valued Gaussian random variables which obeys the following linear model [14],[9]

$$\mathbf{X} = \mathbf{ASB} + \mathbf{W} \tag{9}$$

where $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{B}$ are rectangular $m \times q$, $q \times p$ and $p \times n$ complex matrices, and $\mathbf{W}$ is an $m \times n$ matrix whose n columns are i.i.d. zero mean circular complex Gaussian vectors each with positive definite covariance matrix $\mathbf{R}_w$. We will assume that $n \geq m$. This model is very general and, as will be seen in subsequent sections, covers many signal processing applications.

A few comments about random matrices are now in order. If $\mathbf{Z}$ is an $m \times n$ random matrix the mean, $E[\mathbf{Z}]$, of $\mathbf{Z}$ is defined as the $m \times n$ matrix of means of the elements of $\mathbf{Z}$, and the covariance matrix is defined as the $mn \times mn$ covariance matrix of the $mn \times 1$ vector, vec$[\mathbf{Z}]$, formed by stacking columns of $\mathbf{Z}$. When the columns of $\mathbf{Z}$ are uncorrelated and each have the same $m \times m$ covariance matrix $\mathbf{R}$, the covariance of $\mathbf{Z}$ is block diagonal:

$$\mathrm{cov}[\mathbf{Z}] = \mathbf{R} \otimes \mathbf{I}_n. \tag{10}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. For $p \times q$ matrix $\mathbf{C}$ and $r \times s$ matrix $\mathbf{D}$ the notation $\mathbf{C} \otimes \mathbf{D}$ denotes the kronecker product which is the following $pr \times qs$ matrix:

$$\mathbf{C} \otimes \mathbf{D} = \begin{bmatrix} \mathbf{C}\, d_{11} & \mathbf{C}\, d_{12} & \ldots & \mathbf{C}\, d_{1s} \\ \mathbf{C}\, d_{21} & \mathbf{C}\, d_{22} & \ldots & \mathbf{C}\, d_{2s} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{C}\, d_{r1} & \mathbf{C}\, d_{r2} & \ldots & \mathbf{C}\, d_{rs} \end{bmatrix}. \tag{11}$$

The density function of $\mathbf{X}$ has the form [14]

$$f(\mathbf{X};\theta) = \frac{1}{\pi^{mn}|\mathbf{R}_w|^n} \exp\left(-\mathrm{tr}\left\{[\mathbf{X} - \mathbf{ASB}][\mathbf{X} - \mathbf{ASB}]^H \mathbf{R}_w^{-1}\right\}\right), \tag{12}$$

where $|\mathbf{C}|$ is the determinant and tr$\{\mathbf{D}\}$ is the trace of square matrices $\mathbf{C}$ and $\mathbf{D}$. For convenience we will use the shorthand notation

$$\mathbf{X} \sim \mathcal{N}_{mn}(\mathbf{ASB}, \mathbf{R}_w \otimes \mathbf{I}_n)$$

which is to be read as $\mathbf{X}$ is distributed as an $m \times n$ complex Gaussian random matrix with mean $\mathbf{ASB}$, and covariance $\mathbf{R}_w \otimes \mathbf{I}_n$,

In the examples presented in the next section, several distributions associated with the complex Gaussian distribution will be seen to govern the various test statistics. The complex non-central Chi-Square distribution with p degrees of freedom and vector of non-centrality parameters $(\rho, \underline{d})$ plays a very important role here. This is defined as the distribution of the random variable $\chi^2(\rho, \underline{d}) \overset{\text{def}}{=} \sum_{i=1}^{p} d_i |z_i|^2 + \rho$ where the z_i's are independent univariate complex Gaussian random variables with zero mean and unit variance and where ρ is scalar and $\underline{d}$ is a (row) vector of positive scalars. The complex non-central Chi-square distribution is closely related to the real non-central Chi-square distribution with $2p$ degrees of freedom and non-centrality parameters $(\rho, \text{diag}([\underline{d}, \underline{d}]))$ defined in [9]. The case of $\rho = 0$ and $\underline{d} = [1, \ldots, 1]$ corresponds to the standard (central) complex Chi-square distribution. For derivations and details on this and other related distributions see [14].

5 Temporal Signals in Gaussian Noise

Consider the time sampled superposed signal model

$$x(t_i) = \sum_{j=1}^{p} s_j b_j(t_i) + w(t_i), \quad i = 1, \ldots, n,$$

where here we interpret t_i as time; but it could also be space or other domain. The temporal signal waveforms $\underline{b}_j = [b_j(t_1), \ldots, b_j(t_n)]^T$, $j = 1, \ldots, p$, are assumed to be linearly independent where $p \leq n$. The scalar s_j is a time independent complex gain applied to the j-th signal waveform. The noise $w(t)$ is complex Gaussian with zero mean and correlation function $r_w(t, \tau) = E[w(t)w^*(\tau)]$. By concatenating the samples into a column vector $\underline{x} = [x(t_1), \ldots, x(t_n)]^T$ the above model is equivalent to:

$$\underline{x} = \mathbf{B}\underline{s} + \underline{w}, \tag{13}$$

where $\mathbf{B} = [\underline{b}_1, \ldots, \underline{b}_p]$, $\underline{s} = [s_1, \ldots, s_p]^T$. Therefore the density function (12) applies to the transpose $\underline{x}^T$ with $\mathbf{R}_w = \text{cov}(\underline{w})$, $m = q = 1$, and $\mathbf{A} = 1$.

5.1 Signal Detection: Known Gains

For known gain factors s_i, known signal waveforms $\underline{b}_i$, and known noise covariance $\mathbf{R}_w$, the LRT (5) is the most powerful signal detector for deciding between the simple hypotheses $H : \underline{x} \sim \mathcal{N}_n(0, \mathbf{R}_w)$ versus $K : \underline{x} \sim \mathcal{N}_n(\mathbf{B}\underline{s}, \mathbf{R}_w)$. The LRT has the form

$$L(x) = \exp\left(-2 * \text{Re}\left\{\underline{x}^H \mathbf{R}_w^{-1} \mathbf{B}\underline{s}\right\} + \underline{s}^H \mathbf{B}^H \mathbf{R}_w^{-1} \mathbf{B}\underline{s}\right) \quad \overset{K}{\underset{H}{\gtrless}} \quad \eta. \tag{14}$$

This test is equivalent to a linear detector with critical region $\mathcal{R}_K = \{x : T(x) > \gamma\}$ where

$$T(x) = Re\left\{\underline{x}^H \mathbf{R}_w^{-1} \underline{s}_c\right\}$$

and $\underline{s}_c = \mathbf{B}\underline{s} = \sum_{j=1}^{p} s_j \underline{b}_j$ is the observed compound signal component.

Under both hypotheses H and K the test statistic T is Gaussian distributed with common variance but different means. It is easily shown that the ROC curve is monotonically increasing in the *detectability index* $\rho = \underline{s}_c^H \mathbf{R}_w^{-1} \underline{s}_c$. It is interesting to note that when the noise is white, $\mathbf{R}_w = \sigma^2 \mathbf{I}_n$ and the ROC curve depends on the form of the signals only through the signal-to-noise ratio (SNR) $\rho = \frac{\|\underline{s}_c\|^2}{\sigma^2}$. In this special case the linear detector can be written in the form of a correlator detector

$$T(x) = Re\left\{\sum_{i=1}^{n} s_c^*(t_i)x(t_i)\right\} \quad \overset{K}{\underset{H}{\gtrless}} \quad \gamma$$

where $s_c(t) = \sum_{j=1}^{p} s_j b_j(t)$. When the sampling times t_i are equispaced, e.g. $t_i = i$, the correlator takes the form of a matched filter

$$T(x) = Re\left\{\sum_{i=1}^{n} h(n-i)x(i)\right\} \quad \overset{K}{\underset{H}{\gtrless}} \quad \gamma,$$

where $h(i) = s_c^*(-i)$. Block diagrams for the correlator and matched filter implementations of the LRT are shown in Figs. 3 and 4.

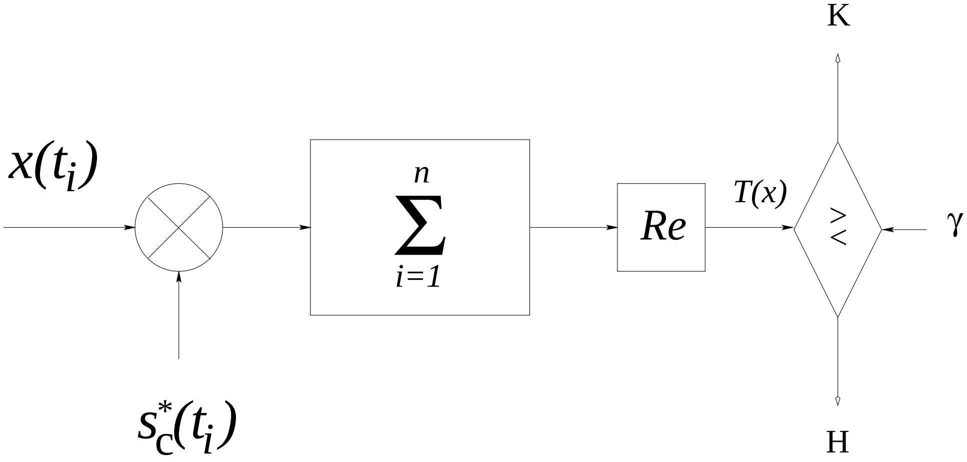

Figure 3: The correlator implementation of the most powerful LRT for signal component $s_c(t_i)$ in additive Gaussian white noise. For non-white noise a prewhitening transformation must be performed on $x(t_i)$ and $s_c(t_i)$ prior to implementation of correlator detector.

5.2 Signal Detection: Unknown Gains

When the gains s_j are unknown the alternative hypothesis K is composite, the critical region $\mathcal{R}_K$ depends on the true gains for $p > 1$, and no most powerfull test for $H : \underline{x} \sim \mathcal{N}_n(0, \mathbf{R}_w)$ versus $K : \underline{x} \sim \mathcal{N}_n(\mathbf{B}\underline{s}, \mathbf{R}_w)$ exists. However, the GLRT (7) can easily be derived by maximizing the likelihood ratio for known gains (14) over $\underline{s}$. Recalling from least squares theory that $\min_{\underline{s}}(\underline{x} -$

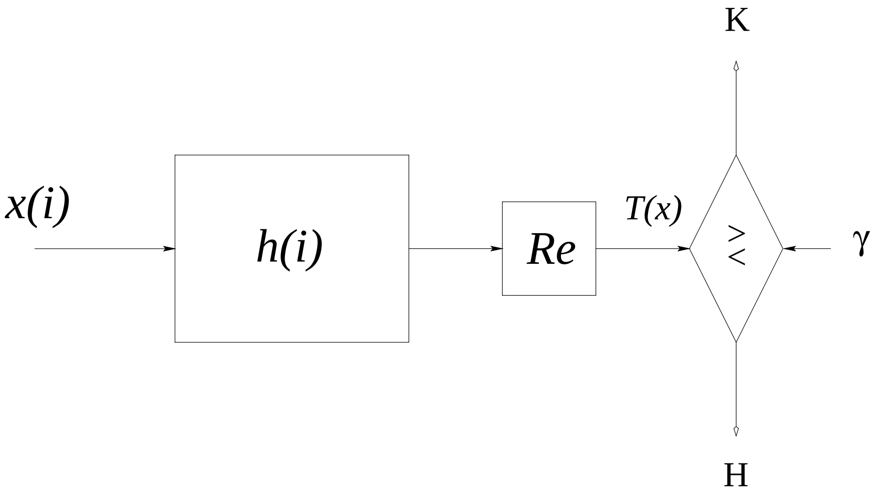

Figure 4: The matched filter implementation of the most powerful LRT for signal component $s_c(i)$ in additive Gaussian white noise. Matched filter impulse response is $h(i) = s_c^*(-i)$. For non-white noise a prewhitening transformation must be performed on $x(i)$ and $s_c(i)$ prior to implementation of matched filter detector.

$\mathbf{B}\underline{s})^H \mathbf{R}_w^{-1}(\underline{x} - \mathbf{B}\underline{s}) = \underline{x}^H \mathbf{R}_w^{-1}\underline{x} - \underline{x}^H \mathbf{R}_w^{-1}\mathbf{B}[\mathbf{B}^H \mathbf{R}_w^{-1}\mathbf{B}]^{-1}\mathbf{B}^H \mathbf{R}_w^{-1}\underline{x}$ the GLRT can be shown to take the form

$$T_g(x) = \underline{x}^H \mathbf{R}_w^{-1}\mathbf{B}[\mathbf{B}^H \mathbf{R}_w^{-1}\mathbf{B}]^{-1}\mathbf{B}^H \mathbf{R}_w^{-1}\underline{x} \quad \overset{K}{\underset{H}{\gtrless}} \quad \gamma.$$

A more intuitive form for the GLRT can be obtained by expressing T_g in terms of the prewhitened observations $\tilde{\underline{x}} = \mathbf{R}_w^{-\frac{1}{2}}\underline{x}$ and prewhitened signal waveform matrix $\tilde{\mathbf{B}} = \mathbf{R}_w^{-\frac{1}{2}}\mathbf{B}$, where $\mathbf{R}_w^{-\frac{1}{2}}$ is the right Cholesky factor of $\mathbf{R}_w^{-1}$

$$T_g(x) = ||\tilde{\mathbf{B}}[\tilde{\mathbf{B}}^H\tilde{\mathbf{B}}]^{-1}\tilde{\mathbf{B}}^H\tilde{\underline{x}}||^2. \tag{15}$$

$\tilde{\mathbf{B}}[\tilde{\mathbf{B}}^H\tilde{\mathbf{B}}]^{-1}\tilde{\mathbf{B}}^H$ is the idempotent $n \times n$ matrix which projects onto column space of the prewhitened signal waveform matrix $\tilde{\mathbf{B}}$ (whitened signal subspace). Thus the GLRT decides that some linear combination of the signal waveforms $\underline{b}_1, \ldots, \underline{b}_p$ is present only if the energy of the component of x lying in the whitened signal subspace is sufficiently large.

Under the null hypothesis the test statistic T_g is distributed as a complex central Chi-Square random variable with p degrees of freedom, while under the alternative hypothesis T_g is non-central Chi-Square with non-centrality parameter vector $(\underline{s}^H\mathbf{B}^H\mathbf{R}_w^{-1}\mathbf{B}\underline{s}, 1)$. The ROC curve is indexed by the number of signals p and the non-centrality parameter but is not expressible in closed form for $p > 1$.

5.3 Signal Detection: Random Gains

In some cases a random Gaussian model for the gains may be more appropriate than the unknown gain model considered above. When the p-dimensional gain vector $\underline{s}$ is multi-variate normal with zero mean and $p \times p$ covariance matrix $\mathbf{R}_s$ the compound signal component $\underline{s}_c = \mathbf{B}\underline{s}$ is an n-dimensional random Gaussian vector with zero mean and rank p covariance matrix $\mathbf{B}\mathbf{R}_s\mathbf{B}^H$. A standard assumption is that the gains and the additive noise are statistically independent. The detection problem can then be stated as testing the two simple hypotheses $H : \underline{x} \sim \mathcal{N}_n(0, \mathbf{R}_w)$ versus $K : \underline{x} \sim \mathcal{N}_n(0, \mathbf{B}\mathbf{R}_s\mathbf{B}^H + \mathbf{R}_w)$. It can be shown that the most powerful LRT has the form

$$T(x) = \sum_{i=1}^{p} \left(\frac{\lambda_i}{1 + \lambda_i} \right) |\underline{v}_i^* \mathbf{R}_w^{-\frac{1}{2}} \underline{x}|^2 \quad \mathop{\gtrless}\limits^{K}_{H} \quad \gamma, \tag{16}$$

where $\{\lambda_i\}_{i=1}^p$ are the non-zero eigenvalues of the matrix $\mathbf{R}_w^{-\frac{1}{2}} \mathbf{B}\mathbf{R}_s\mathbf{B}^H \mathbf{R}_w^{-\frac{H}{2}}$ and $\{\underline{v}_i\}_{i=1}^p$ are the associated eigenvectors. Under H the test statistic $T(x)$ is distributed as complex non-central Chi-square with p degrees of freedom and non-centrality parameter vector $(0, \underline{d}_H)$ where $\underline{d}_H = [\lambda_1/(1 + \lambda_1), \ldots, \lambda_p/(1 + \lambda_p)]$. Under the alternative hypothesis T is also distributed as non-central complex Chi-square, however with non-centrality vector $(0, \underline{d}_K)$ where $\underline{d}_K$ are the non-zero eigenvalues of $\mathbf{B}\mathbf{R}_s\mathbf{B}^H$. The ROC is not available in closed form for $p > 1$.

5.4 Signal Detection: Single Signal

We obtain a unification of the GLRT for unknown gain and the LRT for random gain in the case of a single impinging signal waveform: $\mathbf{B} = \underline{b}_1$, $p = 1$. In this case the test statistic T_g in (15) and T in (16) reduce to the identical form and we get the same detector structure

$$\frac{\left| \underline{x}^H \mathbf{R}_w^{-1} \underline{b}_1 \right|^2}{\underline{b}_1^H \mathbf{R}_w^{-1} \underline{b}_1} \quad \mathop{\gtrless}\limits^{K}_{H} \quad \eta,$$

This establishes that the GLRT is uniformly most powerfull over all values of the gain parameter s_1 for $p = 1$. Note that even though the form of the unknown parameter GLRT and the random parameter LRT are identical for this case, their ROC curves and their thresholds γ will be different since the underlying observation models are not the same. When the noise is white the test simply compares the magnitude squared of the complex correlator output $\sum_{i=1}^n b_1^*(t_i)x(t_i)$ to a threshold γ.

6 Spatio-Temporal Signals

Consider the general spatio-temporal model

$$\underline{x}(t_i) = \sum_{j=1}^{q} \underline{a}_j \sum_{k=1}^{p} s_{jk} b_k(t_i) + \underline{w}(t_i), \quad i = 1, \ldots, n.$$

This model applies to a wide range of applications in narrowband array processing and has been thoroughly studied in the context of signal detection in [14]. The m-element vector $\underline{x}(t_i)$ is a snapshot at time t_i of the m-element array response to p impinging signals arriving from q different directions. The vector $\underline{a}_j$ is a known *steering vector* which is the complex response of the array to signal energy arriving from the j-th direction. From this direction the array receives the superposition $\sum_{k=1}^{p} s_{jk}\underline{b}_k$ of p known time varying signal waveforms $\underline{b}_k = [b_k(t_1), \ldots, b_k(t_n)]^T$, $k = 1, \ldots, p$. The presence of the superposition accounts for both direct and multipath arrivals and allows for more signal sources than directions of arrivals when $p > q$. The complex Gaussian noise vectors $\underline{w}(t_i)$ are spatially correlated with spatial covariance $\mathrm{cov}[\underline{w}(t_i)] = \mathbf{R}_w$ but are temporally uncorrelated $\mathrm{cov}[\underline{w}(t_i), \underline{w}(t_j)] = 0$, $i \neq j$.

By arranging the n column vectors $\{\underline{x}(t_i)\}_{i=1}^{n}$ in an $m \times n$ matrix $\mathbf{X}$ we obtain the equivalent matrix model

$$\mathbf{X} \;=\; \mathbf{ASB}^H + \mathbf{W},$$

where $\mathbf{S} = (s_{ij})$ is a $q \times p$ matrix whose rows are vectors of signal gain factors for each different direction of arrival, $\mathbf{A} = [\underline{a}_1, \ldots, \underline{a}_q]$ is an $m \times q$ matrix whose columns are steering vectors for different directions of arrival, and $\mathbf{B} = [\underline{b}_1, \ldots, \underline{b}_p]^T$ is a $p \times n$ matrix whose rows are different signal waveforms. To avoid singular detection it is assumed that $\mathbf{A}$ is of rank q, $q \leq m$, and that $\mathbf{B}$ is of rank p, $p \leq n$. We consider only a few applications of this model here. For many others see [14].

6.1 Detection: Known Gains and Known Spatial Covariance

First we assume the gain matrix $\mathbf{S}$ and the spatial covariance $\mathbf{R}_w$ are known. This case is only relevant when one knows the direct path and multipath geometry of the propagation medium ($\mathbf{S}$), the spatial distribution of the ambient (possibly coherent) noise ($\mathbf{R}_w$), the q directions of the impinging superposed signals ($\mathbf{A}$), and the p signal waveforms ($\mathbf{B}$). Here the detection problem is stated in terms of the simple hypotheses $H : \mathbf{X} \sim \mathcal{N}_{nm}(0, \mathbf{R}_w \otimes \mathbf{I}_n)$ versus $K : \mathbf{X} \sim \mathcal{N}_{nm}(\mathbf{ASB}, \mathbf{R}_w \otimes \mathbf{I}_n)$. For this case, the LRT (5) is the most powerful test and, using (12), has the form

$$T(x) = \mathrm{Re}\left(\mathrm{tr}\left\{\mathbf{A}^H \mathbf{R}_w^{-1} \mathbf{X} \mathbf{B}^H \mathbf{S}^H\right\}\right) \quad \overset{K}{\underset{H}{\gtrless}} \quad \gamma.$$

Since the test statistic is Gaussian under H and K the ROC curve is of similar form to the ROC for detection of temporal signals with known gains.

Identifying the quantities $\tilde{\mathbf{X}} = \mathbf{R}_w^{-\frac{1}{2}}\mathbf{X}$ and $\tilde{\mathbf{A}} = \mathbf{R}_w^{-\frac{1}{2}}\mathbf{A}$ as the spatially whitened measurement matrix and spatially whitened array response matrix, respectively, the test statistic T can be interpreted as a multivariate spatio-temporal correlator detector. In particular, when there is only one signal impinging on the array from a single direction then $p = q = 1$, $\tilde{\mathbf{A}} = \underline{\tilde{a}}$ a column vector, $\mathbf{B} = \underline{b}^T$ a row vector, $\mathbf{S} = s$ a complex scalar, and the test statistic becomes

$$\begin{aligned} T(x) &= \mathrm{Re}\left\{\underline{\tilde{a}}^H \cdot_s \tilde{\mathbf{X}} \cdot_t \underline{b}^* \, s^*\right\} \\ &= \mathrm{Re}\left\{ s^* \sum_{j=1}^{m} \tilde{a}_j^* \sum_{i=1}^{n} b^*(t_i)\tilde{x}_j(t_i) \right\}. \end{aligned}$$

In the above the multiplication notation $\cdot_s$ and $\cdot_t$ is used to simply emphasize the respective matrix multiplication operations (correlation) which occur over the spatial domain and the time domain. It can be shown that the ROC curve monotonically increases in the detectability index $\rho = n\underline{a}^H \mathbf{R}_w^{-1} \underline{a} \cdot \|s\underline{b}\|^2$.

6.2 Detection: Unknown Gains and Unknown Spatial Covariance

By assuming the gain matrix $\mathbf{S}$ and $\mathbf{R}_w$ to be unknown the detection problem becomes one of testing for noise alone against noise plus p coherent signal waveforms, where the waveforms lie in the subspace formed by all linear combinations of the rows of $\mathbf{B}$ but are otherwise unknown. This gives a composite null and alternative hypothesis for which the generalized likelihood ratio test can be derived by maximizing the known-gain likelihood ratio over the gain matrix $\mathbf{S}$. The result is the GLRT [14]

$$T_g(x) = \frac{\left|\mathbf{A}^H \hat{\hat{\mathbf{R}}}_K^{-1} \mathbf{A}\right|}{\left|\mathbf{A}^H \hat{\mathbf{R}}_H^{-1} \mathbf{A}\right|} \overset{K}{\underset{H}{\gtrless}} \gamma,$$

where $|\cdot|$ denotes the determinant, $\hat{\mathbf{R}}_H = \frac{1}{n}\mathbf{X}\mathbf{X}^H$ is a sample estimate of the spatial covariance matrix using all of the snapshots, and $\hat{\hat{\mathbf{R}}}_K = \frac{1}{n}\mathbf{X}[\mathbf{I}_n - \mathbf{B}^H[\mathbf{B}\mathbf{B}^H]^{-1}\mathbf{B}]\mathbf{X}^H$ is the sample estimate using only those components of the snapshots lying outside of the row space of the signal waveform matrix $\mathbf{B}$. To gain insight into the test statistic T_g consider the asymptotic convergence of T_g as the number of snapshots n goes to infinity. By the strong law $\hat{\hat{\mathbf{R}}}_K$ converges to the covariance matrix of $\mathbf{X}[\mathbf{I}_n - \mathbf{B}^H[\mathbf{B}\mathbf{B}^H]^{-1}\mathbf{B}]$. Since $\mathbf{I}_n - \mathbf{B}^H[\mathbf{B}\mathbf{B}^H]^{-1}\mathbf{B}$ annihilates the signal component $\mathbf{ASB}$, this covariance is the same quantity $\mathbf{R}$, $\mathbf{R} \leq \mathbf{R}_w$, under both H and K. On the other hand, $\hat{\mathbf{R}}_H$ converges to $\mathbf{R}_w$ under H while it converges to $\mathbf{R}_w + \mathbf{ASBB}^H\mathbf{S}^H\mathbf{A}^H$ under K. Hence when strong signals are present T_g tends to take on very large values near the quantity $\left(|\mathbf{A}^H\mathbf{R}^{-1}\mathbf{A}|\right) / \left(|\mathbf{A}^H[\mathbf{R}_w + \mathbf{ASBB}^H\mathbf{S}^H\mathbf{A}^H]^{-1}\mathbf{A}^H|\right) \gg 1$.

The distribution of T_g under H (K) can be derived in terms of the distribution of a sum of central (non-central) complex Beta random variables. See [14] for discussion of performance and algorithms for data recursive computation of T_g. Generalizations of this GLRT exist which incorporate non-zero mean [14, 15].

7 Signal Classification

Typical classification problems arising in signal processing are: classifying an individual signal waveform out of a set of possible linearly independent waveforms, classifying the presence of a particular set of signals as opposed to other sets of signals, classifying among specific linear combinations of signals, and classifying the number of signals present. The problem of classification of the number of signals, also known as the order selection problem, is treated elsewhere in this chapter (Djuric?). While the spatio-temporal model could be treated in analogous fashion, for concreteness we focus on the case of the Gaussian temporal signal model (13).

7.1 Classifying Individual Signals

Here it is of interest to decide which one of the p scaled signal waveforms $s_1\underline{b}_1, \ldots, s_p\underline{b}_p$ are present in the observations $\underline{x} = [x(t_1), \ldots x(t_n)]^T$. Denote by H_k the hypothesis that $\underline{x} = s_k\underline{b_k} + \underline{w}$. Signal classification can then be stated as the problem of testing between the following simple hypotheses

$$H_1 \quad : \quad x = s_1\underline{b}_1 + \underline{w}$$

$$\begin{array}{ccc} \vdots & \vdots & \vdots \\ H_p & : & x = s_p\underline{b}_p + \underline{w} \end{array}$$

For known known gain factors s_k, known signal waveforms $\underline{b}_k$, and known noise covariance $\mathbf{R}_w$, these hypotheses are simple, the density function $f(x|s_k, \underline{b}_k) = \mathcal{N}_n(s_k\underline{b}_k, \mathbf{R}_w)$ under H_k involves no unknown parameters and the maximum likelihood classifier (8) reduces to the decision rule

$$\text{decide } H_j \text{ if and only if } j = \text{argmin}_{k=1,\ldots,p}(\underline{x} - s_k\underline{b}_k)^H\mathbf{R}_w^{-1}(\underline{x} - s_k\underline{b}_k)\ . \qquad (17)$$

Thus the classifier chooses the most likely signal as that signal $s_j\underline{b}_j$ which has minimum normalized distance from the observed waveform $\underline{x}$. The classifier can also be interpreted as a *minimum distance classifier* which chooses the signal which minimizes the Euclidean distance $\|\tilde{\underline{x}} - s_k\tilde{\underline{b}}_k\|$ between the prewhitened signal $\tilde{\underline{b}}_k = \mathbf{R}_w^{-\frac{1}{2}}\underline{b}_k$ and the prewhitened measurement $\tilde{\underline{x}} = \mathbf{R}_w^{-\frac{1}{2}}\underline{x}$.

Written in the minimum normalized distance form, the ML classifier appears to involve nonlinear statistics. However, an obvious simplification of (17) reveals that the ML classifier actually only requires computing linear functions of $\underline{x}$

$$\text{decide } H_j \text{ if and only if } j = \text{argmax}_{k=1,\ldots,p}\left\{\text{Re}\left(\underline{x}^H\mathbf{R}_w^{-1}\underline{b}_k\ s_k\right) - \tfrac{1}{2}|s_k|^2\ \underline{b}_k^H\mathbf{R}_w^{-1}\underline{b}_k\right\}\ .$$

Note that this linear reduction only occurs when the covariances $\mathbf{R}_w$ are identical under each H_k, $k = 1, \ldots, p$. In this case the ML classifier can be implemented using prewhitening filters followed by a bank of correlators or matched filters, an offset adjustment, and a maximum selector (Fig. 5).

An additional simplification occurs when the noise is white, $\mathbf{R}_w = \mathbf{I}_n$, and all signal energies $|s_k|^2\|\underline{b}_k^H\|^2$ are identical: the classifier chooses the most likely signal as that signal $b_j(t_i)s_j$ which is maximally correlated with the measurement x:

$$\text{decide } H_j \text{ if and only if } j = \text{argmax}_{k=1,\ldots,p}\text{Re}\left(s_k \textstyle\sum_{i=1}^n b_k^*(t_i)x(t_i)\right).$$

The decision regions $\mathcal{R}_{H_k} = \{x : \text{decide } H_k\}$ induced by (17) are piecewise linear regions, known as Voronoi cells $\mathcal{V}_k$, centered at each of the prewhitened signals $s_k\tilde{\underline{b}}_k$. The misclassification error probabilities $P_{M_k} = 1 - P(x \in \mathcal{R}_{H_k}|H_k) = 1 - \int_{x\in\mathcal{V}_k} f(x|H_k)dx$ must generally be computed by integrating complex multivariate Gaussian densities $f(x|H_k) = \mathcal{N}_n(s_k\underline{b}_k, \mathbf{R}_w)$ over these regions. In the case of orthogonal signals $\underline{b}_i\mathbf{R}_w^{-1}\underline{b}_j = 0$, $i \neq j$, this integration reduces to a single integral of a univariate $\mathcal{N}_1(\rho_k, \rho_k)$ density function times the product of $p-1$ univariate $\mathcal{N}_1(0, \rho_i)$ cumulative distribution functions, $i = 1, \ldots, p, i \neq k$, where $\rho_k = \underline{b}_k^H\mathbf{R}_w^{-1}\underline{b}_k$. Even for this case no general closed form expressions for P_{M_k} is available. However, analytical lower bounds on P_{M_k} and on average missclassification probability $\frac{1}{p}\sum_{k=1}^p P_{M_k}$ can be used to qualitatively assess classifer performance [13].

7.2 Classifying Presence of Multiple Signals

We next treat the problem where the signal component of the observation is the linear combination of one of J hypothesized subsets $\mathcal{S}_k$, $k = 1, \ldots, J$, of the signal waveforms $\underline{b}_1, \ldots, \underline{b}_p$. Assume that subset $\mathcal{S}_k$ contains p_k signals and that the $\mathcal{S}_k$, $k = 1, \ldots, J$, are disjoint, i.e. they do not contain

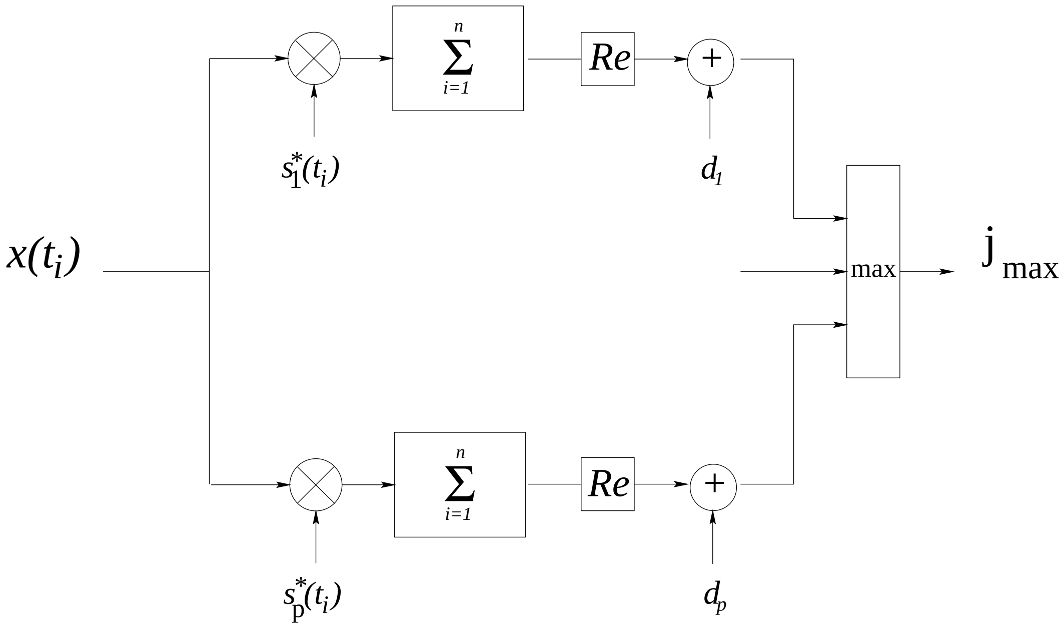

Figure 5: The ML classifier for classifying presence of one of p signals $s_j(t_i) \overset{\text{def}}{=} s_j \underline{b}_j(t_i)$, $j = 1, \ldots, p$, under additive Gaussian white noise. $d_j = -\frac{1}{2}|s_j|^2 \, \|\underline{b}_j\|^2$ and j_{max} is index of correlator output which is maximum. For non-white noise a prewhitening transformation must be performed on $x(t_i)$ and the $b_j(t_i)$'s prior to implementation of ML classifier.

any signals in common. Define the $n \times p_k$ matrix $\mathbf{B}_k$ whose columns are formed from the subset $\mathcal{S}_k$. We can now state the classification problem as testing between the J composite hypotheses

$$\begin{array}{ccl} H_1 & : & \underline{x} = \mathbf{B}_1 \underline{s}_1 + \underline{w}, \quad \underline{s}_1 \in \mathbb{C}^{p_1} \\ \vdots & \vdots & \vdots \\ H_J & : & \underline{x} = \mathbf{B}_J \underline{s}_J + \underline{w}, \quad \underline{s}_J \in \mathbb{C}^{p_J} \end{array}$$

where $\underline{s}_k$ is a column vector of p_k unknown complex gains.

The density function under H_k, $f(x|\underline{s}_k, \mathbf{B}_k) = \mathcal{N}_n(\mathbf{B}_k \underline{s}_k, \mathbf{R}_w)$, is a function of unknown parameters $\underline{s}_k$ and therefore the ML classifier (8) involves finding the largest among maximized likelihoods $\max_{\underline{s}_k} f(x|\underline{s}_k, \mathbf{B}_k)$, $k = 1, \ldots, J$. This yields the following form for the ML classifier:

$$\text{decide } H_j \text{ if and only if } j = \text{argmin}_{k=1,\ldots,J} (\underline{x} - \mathbf{B}_k \underline{\hat{s}}_k)^H \mathbf{R}_w^{-1} (\underline{x} - \mathbf{B}_k \underline{\hat{s}}_k),$$

where $\underline{\hat{s}}_k = \left[\mathbf{B}_k^H \mathbf{R}_w^{-1} \mathbf{B}_k \right]^{-1} \mathbf{B}_k^H \mathbf{R}_w^{-1} \underline{x}$ is the maximum likelihood gain vector estimate. The decision regions are once again piecewise linear but with Voronoi cells having centers at the the least squares estimates of the hypothesized signal components $\mathbf{B}_k \underline{\hat{s}}_k$, $k = 1, \ldots, J$.

Similarly to the case of non-composite hypotheses considered in the previous subsection, a simplification of (18) is possible

$$\text{decide } H_j \text{ if and only if } j = \text{argmax}_{k=1,\ldots,J} \underline{x}^H \mathbf{R}_w^{-1} \mathbf{B}_k [\mathbf{B}_k^H \mathbf{R}_w^{-1} \mathbf{B}_k]^{-1} \mathbf{B}_k^H \mathbf{R}_w^{-1} \underline{x}$$

Defining the prewhitened versions $\underline{\tilde{x}} = \mathbf{R}_w^{-\frac{1}{2}} \underline{x}$ and $\tilde{\mathbf{B}}_k = \mathbf{R}_w^{-\frac{1}{2}} \mathbf{B}_k$ of the observations and the k-th signal matrix, the ML classifier is seen to decide that the linear combination of the p_j signals in H_j is present when the length $\|\tilde{\mathbf{B}}_j [\tilde{\mathbf{B}}_j^H \tilde{\mathbf{B}}_j]^{-1} \tilde{\mathbf{B}}_j^H] \, \underline{\tilde{x}}\|$ of the projection of $\underline{\tilde{x}}$ onto the j-th signal space (colspan$\{\tilde{\mathbf{B}}_j\}$) is greatest. This classifer can be implemented as a bank of p *adaptive* matched filters each matched to one of the least squares estimates $\tilde{\mathbf{B}}_k \underline{\hat{s}}_k$, $k = 1, \ldots, p$, of the prewhitened signal component. Under any H_i the quantities $\underline{x}^H \mathbf{R}_w^{-1} \mathbf{B}_k [\mathbf{B}_k^H \mathbf{R}_w^{-1} \mathbf{B}_k]^{-1} \mathbf{R}_w^{-1} \underline{x}$, $k = 1, \ldots J$, are distributed as complex non-central Chi-square with p_k degrees of freedom. For the special case of orthogonal prewhitened signals $\underline{b}_i \mathbf{R}_w^{-1} \underline{b}_j = 0$, $i \neq j$, these variables are also statistically independent and P_{M_i} can be computed as a one dimensional integral of a univariate non-central Chi-square density times the product of $J-1$ univariate non-central Chi-square cumulative distribution functions.

References

[1] E. L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York, 1959.

[2] T. S. Ferguson, *Mathematical Statistics - A Decision Theoretic Approach*, Academic Press, Orlando FL, 1967.

[3] D. Middleton, *An Introduction to Statistical Communication Theory*, Peninsula Publishing Co, Los Altos CA (Reprint of 1960 McGraw-Hill edition), 1987.

[4] W. Davenport and W. Root, *An introduction to the theory of random signals and noise*, IEEE Press, New York (reprint of 1958 McGraw-Hill edition), 1987.

[5] H. L. Van-Trees, *Detection, Estimation, and Modulation Theory: Part I*, Wiley, New York, 1968.

[6] D. Blackwell and M. A. Girshik, *Theory of Games and Statistical Decisions*, Wiley, New York, 1954.

[7] C. Helstrom, *Elements of signal detection and estimation*, Prentice-Hall, Englewood Cliffs, 1995.

[8] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Addison-Wesley, Reading, MA, 1991.

[9] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.

[10] D. Siegmund, *Sequential analysis: tests and confidence intervals*, Springer-Verlag, New York, 1985.

[11] B. Baygun and A. O. Hero, "Optimal simultaneous detection and estimation under a false alarm constraint," *IEEE Trans. on Inform. Theory*, vol. 41, no. 3, pp. 688–703, 1995.

[12] S. Kassam and J. Thomas, *Nonparametric detection - theory and applications*, Dowden, Hutchinson and Ross, 1980.

[13] K. Fukunaga, *Statistical Pattern Recognition (2nd Ed)*, Academic Press, San Diego CA, 1990.

[14] E. J. Kelly and K. M. Forsythe, "Adaptive detection and parameter estimation for multidimensional signal models," Technical Report 848, M.I.T. Lincoln Laboratory, April, 1989.

[15] T. Kariya and B. K. Sinha, *Robustness of Statistical Tests*, Academic Press, San Diego, 1989.